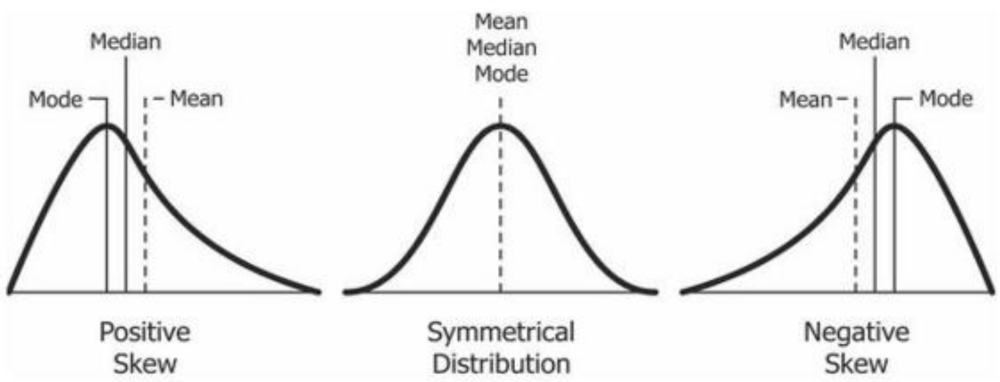
Basic Statistics

	Ungrouped Data	Grouped Data: Discrete Data	Grouped Data: Continuous Data
Arithmetic Mean Or **Mean** Or **Expectation**	→ If $x_1, x_2, x_3 \dots, x_n$ are n observation then the mean is define by, $\bar{x} = \frac{\sum x_i}{n}$	→ If $x_1, x_2, x_3 \dots, x_n$ are n observation and $f_1, f_2, f_3, \dots f_n$ be the corresponding frequencies then the mean is define by, $\bar{x} = \frac{\sum (x_i f_i)}{N}$ Where $N = \sum f_i$	
	Short-cut Method: $\bar{x} = A + \frac{\sum d_i}{n}$ Where $d_i = x_i - A$ and A is assumed mean	Short-cut Method: $\bar{x} = A + \frac{\sum (f_i d_i)}{N}$ Where $d_i = x_i - A$ and A is assumed mean	Short-cut Method: $\bar{x} = A + \frac{\sum (f_i d_i)}{N} \times h$ Where $d_i = \frac{x_i - A}{h}$ and A is assumed mean. h is length of class-interval
Median	→ First arrange all the data in ascending or descending order → If $\boldsymbol{n}$ **is odd** then median is given by, $M = \left(\frac{n+1}{2}\right)^{th}$ *observation* → If $\boldsymbol{n}$ **is even** then median is, $M = \frac{1}{2}\left[\left(\frac{n}{2}\right)^{th} obser. + \left(\frac{n}{2}+1\right)^{th} obser.\right]$	→ First arrange the data in ascending or descending form and find the cumulative frequency table → If $\boldsymbol{N}$ **is odd** then median is given by, $M = \left(\frac{N+1}{2}\right)^{th}$ *observation* → If $\boldsymbol{N}$ **is even** then median is, $M = \frac{1}{2}\left[\left(\frac{N}{2}\right)^{th} obser. + \left(\frac{N}{2}+1\right)^{th} obser.\right]$	$M = l + \frac{\left(\frac{N}{2} - cf\right)}{f} \times h$ Where, l = lower limit of median class h = length of median class f = frequency of median class cf = cumulative frequency of the class preceding to median class **Median class** is the class whose cumulative frequency is just greater than or equal to $\frac{N}{2}$
Mode	→ Mode is the value of observation which is repeated most often in ungroup data → If there are two or more observation in the data that are repeated most often then each such member is a mode.	→ Mode for discrete group data is the value of an observation with highest frequency	$M = l + \left(\frac{f_m - f_1}{2f_m - f_2 - f_1}\right) \times h$ Where, l = lower limit of model class f_1 = frequency of class preceding to model class f_m = frequency of model class f_2 = frequency of class succeeding to model class h = length of the model class **Model class** is the class with maximum frequency
Harmonic Mean	$H.M. = \frac{n}{\sum\left(\frac{1}{x_i}\right)}$	$H.M. = \frac{\sum f_i}{\sum\left(\frac{f_i}{x_i}\right)}$	
Geometric Mean	$\log(G.M.) = \frac{\sum (\log x_i)}{n}$	$\log(G.M.) = \frac{\sum f_i (\log x_i)}{N}$ Note : $H.M. \leq G.M. \leq A.M.$	
Quartiles	→ It is the value on the data which divide the entire data into **four** equal parts	$Q_k = l + \left(\frac{\frac{kN}{4} - cf}{f}\right) \times h \quad where\ k = 1,2,3$	
Deciles	→ It is the value on the data which divide the entire data into **ten** equal parts	$D_k = l + \left(\frac{\frac{kN}{10} - cf}{f}\right) \times h \quad where\ k = 1,2,3, \dots, 9$	
Percentiles	→ It is the value on the data which divide the entire data into **hundred** equal parts	$P_k = l + \left(\frac{\frac{kN}{100} - cf}{f}\right) \times h \quad where\ k = 1,2,3, \dots, 99$	
Mean Deviation from mean $\bar{x}$	$M.D.(\bar{x}) = \frac{\sum \lvert x_i - \bar{x} \rvert}{n}$	$M.D.(\bar{x}) = \frac{\sum f_i \lvert x_i - \bar{x} \rvert}{N}$	

<table>
<tr><td>Mean Deviation from median M</td><td>$$M.D.(M) = \frac{\sum |x_i - M|}{n}$$</td><td colspan="2">$$M.D.(M) = \frac{\sum f_i |x_i - M|}{N}$$</td></tr>
<tr><td rowspan="2">Variance ($\boldsymbol{\sigma^2}$)</td><td>$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2}{n}$$
Or
When value of mean is fraction then,
$$\sigma^2 = \frac{\sum x_i^2}{n} - (\overline{x})^2$$</td><td colspan="2">$$\sigma^2 = \frac{\sum f_i (x_i - \overline{x})^2}{N}$$
Or
When value of mean is fraction then,
$$\sigma^2 = \frac{\sum f_i x_i^2}{N} - (\overline{x})^2$$</td></tr>
<tr><td>Short-cut method:
$\sigma^2 = \frac{1}{n^2}[n \sum d_i^2 - (\sum d_i)^2]$
Where $d_i = x_i - A$
A is assume value from observation</td><td>Short-cut method:
$\sigma^2 = \frac{1}{N^2}[N \sum f_i d_i^2 - (\sum f_i d_i)^2]$
Where $d_i = x_i - A$
A is assume value from observation</td><td>Short-cut method:
$\sigma^2 = \frac{h^2}{N^2}[N \sum f_i d_i^2 - (\sum f_i d_i)^2]$
Where $d_i = \frac{x_i - A}{h}$
A is assume value from observation
h is a length of class interval.</td></tr>
<tr><td>Standard Deviation ($\boldsymbol{\sigma}$)</td><td colspan="3">The standard Deviation (S.D.) $\sigma = \sqrt{variance}$</td></tr>
<tr><td>Coefficient of Variation (C.V.)</td><td colspan="3">$$\boldsymbol{C.V. = \frac{\sigma}{\overline{x}} \times 100}$$
→ If the data has high value of C.V., then it is less consistent
→ If the data has less value of C.V. then it is more consistent</td></tr>
<tr><td>Coefficient of Mean Deviation</td><td colspan="3">Coefficient of Mean Deviation
$$\boldsymbol{= \frac{M.D.}{\overline{x}} \times 100}$$</td></tr>
<tr><td>Central Moments
Or
Moment about mean</td><td>→ The r^{th} moment of n observation of a given data about mean is,
$$\mu_r = \frac{\sum (x_i - \overline{x})^r}{n}$$</td><td colspan="2">→ The r^{th} moment of a given grouped data about mean is,
$$\mu_r = \frac{\sum f_i (x_i - \overline{x})^r}{N}$$</td></tr>
<tr><td></td><td colspan="3">• The first moment about the mean is always zero, ($\boldsymbol{\mu_1 = 0}$)
• The second moment about the mean measures the value of variance, ($\boldsymbol{\mu_2 = \sigma^2}$)
• The third moment about the mean measures the skewness.
• The fourth moment about the mean measures the kurtosis.</td></tr>
<tr><td>Raw Moments
Or
Moment about arbitrary origin A</td><td>→ The r^{th} moment of n observation of a given data about any assume arbitrary value A is,
$$\mu'_r = \frac{\sum (x_i - A)^r}{n}$$</td><td colspan="2">→ The r^{th} moment of a given grouped data about any assume arbitrary value A is,
$$\mu'_r = \frac{\sum f_i (x_i - A)^r}{N}$$</td></tr>
<tr><td colspan="4">Relation between central moment ($\boldsymbol{\mu_r}$) and raw moments ($\boldsymbol{\mu'_r}$):
• $\mu_1 = 0$
• $\mu_2 = \mu'_2 - (\mu'_1)^2$
• $\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3$
• $\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 + (\mu'_1)^4$
Similarly,
• $\mu'_1 = \overline{x} - A$
• $\mu'_2 = \mu_2 + (\mu'_1)^2$
• $\mu'_3 = \mu_3 + 3\mu_2\mu'_1 + (\mu'_1)^3$
• $\mu'_4 = \mu_4 + 4\mu_3\mu'_1 + 6\mu_2(\mu'_1)^2 + (\mu'_1)^4$</td></tr>
<tr><td>Skewness</td><td colspan="3">It is a measure that belongs to the extent of symmetry or asymmetry in a distribution
• Co-efficient of skewness
$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$
• Karl Pearson’s coefficient of skewness
$$\boldsymbol{S_k = \frac{Mean - Mode}{S.D.} = \frac{3(Mean - Median)}{S.D.}}$$</td></tr>
</table>



1) **Symmetrical Distribution:**
   - A distribution is said to be symmetrical if mean = median = mode
   - The graph of symmetrical distribution is perfectly bell shaped curve (Normal curve).

2) **Positively Skewed Distribution:**
   - A distribution is said to be positively skewed distribution if mean > median > mode

3) **Negatively Skewed Distribution:**
   - A distribution is said to be negatively skewed distribution if mean < median < mode

- If $\boldsymbol{\mu_3 = 0}$ then the distribution is said to be symmetrical ($S_k = 0$)
- If $\boldsymbol{\mu_3 > 0}$ then the distribution is said to be positively skewed ($S_k > 0$)
- If $\boldsymbol{\mu_3 < 0}$ then the distribution is said to be negatively skewed ($S_k < 0$)

Note:

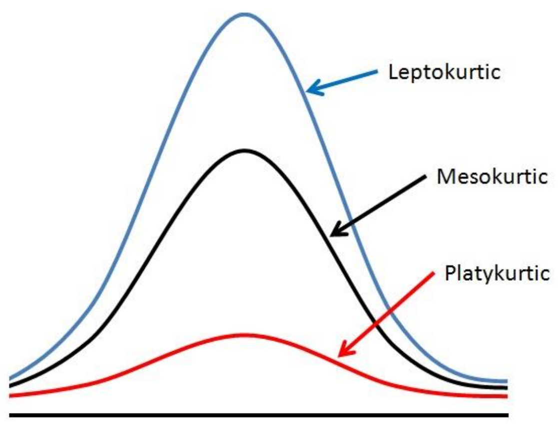
- **The value of Karl Pearson's coefficient of skewness $\boldsymbol{S_k} \in [-\mathbf{1}, \mathbf{1}]$**
- **Use absolute value of $\boldsymbol{S_k}$ to compare the two data. Higher the absolute value of $\boldsymbol{S_k}$ means that data is more skewed.**

Kurtosis

It measures the degree of peakedness of graph of distribution

- **Coefficient of Kurtosis $\boldsymbol{\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}}$**

1) **Mesokurtic:**
   - The curve or graph which is neither flat nor peaked is called Normal curve or Mesokurtic ($\boldsymbol{\beta_2 = 3}$)
2) **Leptokurtic:**
   - The curve which is more peaked then normal curve is called Leptokurtic ($\boldsymbol{\beta_2 > 3}$)
3) **Platykurtic:**
   - The curve which is less peaked then normal curve is called Platykurtic ($\boldsymbol{\beta_2 < 3}$)



Curve Fitting

- Curve fitting is a process of finding the equation of 'best fit' curve for a given set of data.
- It is the representation of the relationship between two variables in form of algebraic equation
- Method of Least square is use to find the best approximate curve for a given set of data

Fitting of Straight line (Linear equation of First Degree)

Let the given set of $'n'$ data is $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$. Our aim is to find the approximate algebraic relationship between x and y which satisfy the all given data points.

(1) Fitting of $\boldsymbol{y = a + bx}$

→ we want to find best approximate equation of straight line of the form

$$y = a + bx$$

→ To find the value of two unknowns a & b we construct two Normal equations of given straight line in terms of a & b. By solving these normal equations we get the value of a & b

→ The Normal equations are,

$$\sum y = an + b\sum x$$
$$\sum xy = a\sum x + b\sum x^2$$

Fitting of Parabola (Quadratic equation of Second degree)

(1) Fitting of $\boldsymbol{y = a + bx + cx^2}$

→ Here we have three unknowns so we can construct three normal equations as,

→ The normal equations are,

$$\sum y = an + b\sum x + c\sum x^2$$
$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$
$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

By solving above three equations we get the value of three unknowns a, b & c.

(2) Fitting of $y = ax + bx^2$

- → Here we have two unknowns so we can construct two normal equations as,
- → The normal equations are,

$$\sum xy = a\sum x^2 + b\sum x^3$$
$$\sum x^2 y = a\sum x^3 + b\sum x^4$$

By solving above two equations we get the value of three unknowns a, b

(3) Fitting of $y = a + bx^2$

- → Here we have two unknowns so we can construct two normal equations as,
- → The normal equations are,

$$\sum y = an + b\sum x^2$$
$$\sum x^2 y = a\sum x^2 + b\sum x^4$$

By solving above two equations we get the value of three unknowns a, b

Fitting of Exponential curve

(1) Fitting of $y = ae^{bx}$

- → We can convert a given exponential curve into linear equation by taking natural logarithm. So,

$$\ln(y) = \ln(a) + bx$$

Now let's say, $Y = \ln(y)$, $A = \ln(a)$, $B = b$, $X = x$. So above equation become linear as,

$$Y = A + BX$$

- → Now in the new linear equation, we have two unknowns so we can construct two normal equations as,
- → The normal equations are,

$$\sum Y = An + B\sum X$$
$$\sum XY = a\sum X + B\sum X^2$$

Where $\boldsymbol{X = x \ and \ Y = \ln(y)}$

- → By solving above two equations we get the value of two unknowns $A \ and \ B$
- → Now to find the value of $a \ and \ b$ we use the relation

$$\boldsymbol{a = e^A \quad and \quad b = B}$$

(2) Fitting of $y = ab^x$

- → We can convert a given exponential curve into linear equation by taking natural logarithm. So,

$$\ln(y) = \ln(a) + x\ln(b)$$

Now let's say, $Y = \ln(y)$, $A = \ln(a)$, $B = \ln(b)$, $X = x$. So above equation become linear as,

$$Y = A + BX$$

- → Now in the new linear equation, we have two unknown so we can construct two normal equations as,
- → The normal equations are,

$$\sum Y = An + B\sum X$$
$$\sum XY = a\sum X + B\sum X^2$$

Where $\boldsymbol{X = x \ and \ Y = \ln(y)}$

- → By solving above two equations we get the value of two unknowns $A \ and \ B$
- → Now to find the value of $a \ and \ b$ we use the relation

$$\boldsymbol{a = e^A \quad and \quad b = e^B}$$

(3) Fitting of $y = ax^b$

- → We can convert a given exponential curve into linear equation by taking natural logarithm. So,

$$\ln(y) = \ln(a) + b\ln(x)$$

Now let's say, $Y = \ln(y)$, $A = \ln(a)$, $B = b$, $X = \ln(x)$. So above equation become linear as,

$$Y = A + BX$$

- → Now in the new linear equation, we have two unknown so we can construct two normal equations as,
- → The normal equations are,

$$\sum Y = An + B\sum X$$
$$\sum XY = a\sum X + B\sum X^2$$

Where $\boldsymbol{X = \ln(x) \ and \ Y = \ln(y)}$

- → By solving above two equations we get the value of two unknowns $A \ and \ B$
- → Now to find the value of $a \ and \ b$ we use the relation

$$\boldsymbol{a = e^A \quad and \quad b = B}$$

Coefficient of Correlation (r)

The correlation between any two variables is a measurement of how the value of one variable is changed when the value of other variable is changed. There are mainly two types of correlation exist.

(1) Positive Correlation ⟶ The change in the value of both variables is in the same direction

- If value of one variable is increase, then the value of other is also increase
- If value of one variable is decrease, then the value of other is also decrease

(2) Negative Correlation ⟶ The change in the value of both variables is in the opposite direction

- If value of one variable is increase, then the value of other is decrease
- If value of one variable is decrease, then the value of other is increase

❖ Karl Pearson's Coefficient of Correlation

It is measures of correlation between two variables $x \ and \ y$ and it is denoted by r

$$r = \frac{cov(x, y)}{\sigma_x \, \sigma_y} = \frac{\frac{1}{n}\sum[(x-\overline{x})(y-\overline{y})]}{\sqrt{\frac{1}{n}\sum(x-\overline{x})^2}\,\sqrt{\frac{1}{n}\sum(y-\overline{y})^2}}$$

→ When $\overline{x}$ and $\overline{y}$ are integer then,

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

→ When the value of both the variables x and y are small or $\overline{x}$ and $\overline{y}$ are not integer then,

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

→ When the value of both the variables x and y are large then,

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

Note:

(1) If $r = 1$ then there exist **perfect positive** correlation between $x \ and \ y$

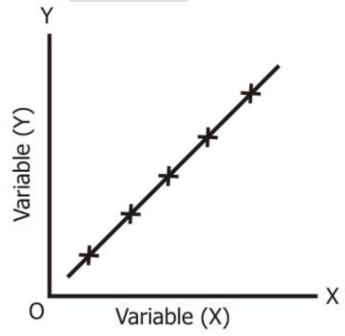
(2) If $r = -1$ then there exist **perfect negative** correlation between $x \ and \ y$

(3) If $0 < r < 1$ then there exist **partial positive** correlation between $x \ and \ y$

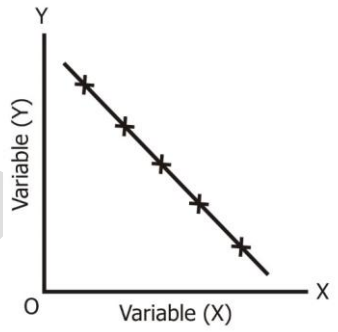
(4) If $-1 < r < 0$ then there exist **partial negative** correlation between $x \ and \ y$

(5) If $r = 0$ then there is **No** correlation between $x \ and \ y$

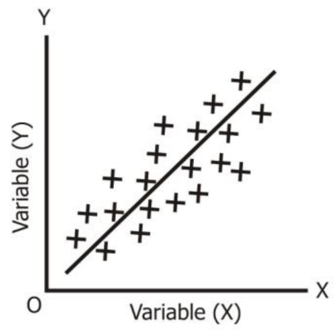
(6) Hence, the value of r is always between $-1 \le r \le 1$



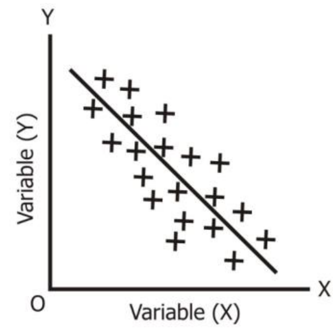
(Perfect Positive)



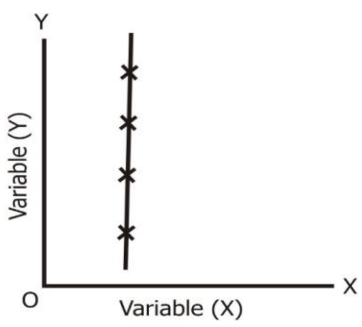
(Perfect Negative)

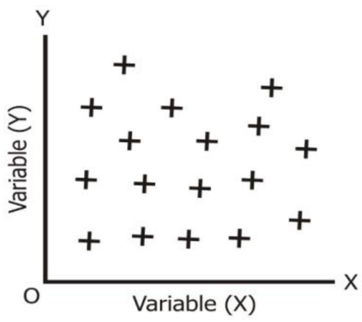


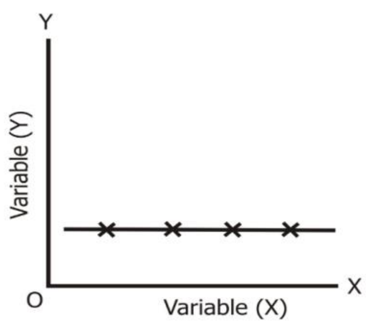
(Partial Positive)



(Partial Negative)







(No Correlation)

Spearman's Rank Correlation

- In this method we assign a rank to each observations either in ascending or in descending order. The spearman's rank correlation coefficient between two variables x and y are given by
(when observation is not repeated)

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

Where $\boldsymbol{d = R_x - R_y}$ and n is total no. of observations

- The spearman's rank correlation coefficient between two variables x and y are given by
(when observation is repeated then rank is distributed equally among the equal observations)

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

Where $\boldsymbol{m}$ denote the no. of repetitions of that particular observation.

Regression

- Regression analysis is used to find the best approximate value of one variable when the value of other variable is known.

- For that we can find the two best approximate linear curve of the form $\boldsymbol{y = a + bx}$ and $\boldsymbol{x = a + by}$ to the given set of $'n'$ data points $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$.

x	x_1	x_2	x_3	………………...	x_n
y	y_1	y_2	y_3	………………...	y_n

→ The linear curve of the form
$\boldsymbol{y = a + bx} \longrightarrow$ The regression line of $\boldsymbol{y\ on\ x}$ [It is also denoted by $y = f(x)$]
This regression line gives the best approximate value of variable $\boldsymbol{y}$ when the value of $\boldsymbol{x}$ is known.

→ The linear curve of the form
$\boldsymbol{x = a + by} \longrightarrow$ The regression line of $\boldsymbol{x\ on\ y}$ [It is also denoted by $x = f(y)$]
This regression line gives the best approximate value of variable $\boldsymbol{x}$ when the value of $\boldsymbol{y}$ is known.

(1) Regression line of $y\ on\ x$ or $[y = a + bx]$

→ The equation of regression line of $y\ on\ x$ is given by,

$$\boldsymbol{y - \overline{y} = b_{yx}\,(x - \overline{x}\,)}$$

Where,
$\overline{x} = mean\ value\ of\ x$
$\overline{y} = mean\ value\ of\ y$
$b_{yx} = Regression\ coefficient\ of\ y\ on\ x$

→ The value of regression coefficient of $y\ on\ x$ is given by,

$$\boldsymbol{b_{yx} = r\frac{\sigma_y}{\sigma_x}}$$

We have three formula to calculate the value of b_{yx}

→ When $\overline{x}$ and $\overline{y}$ are integer then,

$$\boldsymbol{b_{yx} = \frac{\sum[(x-\overline{x}\,)(y-\overline{y}\,)]}{\sum(x-\overline{x}\,)^2}}$$

→ When the value of both the variables x and y are small or $\overline{x}$ and $\overline{y}$ are not integer then,

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

→ When the value of both the variables x and y are large then,

$$\boldsymbol{b_{yx} = \frac{n\sum(d_x d_y) - (\sum d_x)(\sum d_y)}{n\sum d_x^2 - (\sum d_x)^2}}$$

(2) **Regression line of x on y or $[x = a + by]$**

→ The equation of regression line of x on y is given by

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

Where,
$\overline{x} = mean\ value\ of\ x$
$\overline{y} = mean\ value\ of\ y$
$b_{xy} = Regression\ coefficient\ of\ x\ on\ y$

→ The value of regression coefficient of x on y is given by,

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

We have three formula to calculate the value of b_{xy}

→ When $\overline{x}$ and $\overline{y}$ are integer then,

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

→ When the value of both the variables x and y are small or $\overline{x}$ and $\overline{y}$ are not integer then,

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

→ When the value of both the variables x and y are large then,

$$b_{xy} = \frac{n\sum(d_x d_y) - (\sum d_x)(\sum d_y)}{n\sum d_y^2 - (\sum d_y)^2}$$

Note:

(1) The value of co-efficient of correlation r is given by, $r = \pm\sqrt{b_{yx} \times b_{xy}}$ such that $-1 \le r \le 1$.

(2) If value of one regression coefficient is greater than one then other must be less than one. (converse may be not true)

(3) The equation of both regression lines intersect at their mean or at point $(\overline{x}, \overline{y})$.

(4) The sign of both regression coefficients and the sign of correlation coefficient are same.
If $b_{yx} > 0 \Leftrightarrow b_{xy} > 0 \Leftrightarrow r > 0$ *and* If $b_{yx} < 0 \Leftrightarrow b_{xy} < 0 \Leftrightarrow r < 0$

(5) If any equation is convert in the form $y = a + bx$ then b is represent the regression coefficient of y on x. So $b = b_{yx}$

(6) If any equation is convert in the form $x = a + by$ then b is represent the regression coefficient of x on y. So $b = b_{xy}$

(7) If $r = 0$ then both regression lines are perpendicular.

(8) If $r = \pm 1$ then both regression lines are coincide.