Chapter Three

Averages and Measures of Central Tendency

An average is a value which is typical or representative of a set of data. Such typical values tend to lie centrally when the data are arranged according to magnitude. Therefore, averages are also called measures of central tendency.

1. *Arithmetic Mean*:

This is the most common type of average, briefly called as the "*Mean*". It is the sum of the observations divided by their number.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Example:

Consider the following 20 observations ordered from the smallest to the largest, each one representing the lifetime in hours of a certain type of lamp.

612, 623, 666, 744, 883, 898, 964, 970, 983, 1003, 1016, 1022, 1029, 1058, 1085, 1088, 1122, 1135, 1197, 1201.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{612 + 623 + \ldots + 1201}{20} = \frac{19299}{20} = 964.95$$

Grouped data:

When the frequency of some of the observations is greater than 1, computation may be simplified by using frequency grouping. If f_i is the frequency of any value x_i, then the mean for grouped data can be written as:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} \quad \text{, where } \sum f_i = n$$

Example:

1. Find the arithmetic mean for the following data.
2. Construct a frequency distribution for the data and find the arithmetic mean for the grouped data.

37, 43, 42, 46, 37, 44, 38, 39, 37, 42, 38, 45, 38, 48, 43

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{37 + 43 + \ldots + 43}{15} = \frac{617}{15} = 41.13$$

Range = 48 – 37 = 11

Class width = 11/4 = 2.75 ≈ 3

Lower class limit = 37 Upper class limit = 39

Class	f_i	x_i	$f_i x_i$	d_i	$f_i d_i$
37 – 39	7	38	266	0	0
40 – 42	2	41	82	3	6
43 – 45	4	44	176	6	24
46 – 48	2	47	94	9	18
	15		618		48

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{618}{15} = 41.2$$

Coding Method:

A further saving in effort is effected by reducing all observations by a constant value or by dividing them by a factor such as 10, this is known as coding. Hence if A is an arbitrary constant, $d_i = x_i - A$,

Then: $\bar{x} = A + \frac{\sum d_i}{n}$ for ungrouped data

Let $A = 37$, then d_i becomes: 0, 6, 5, 9, 0, 7, 1, 2, 0, 5, 1, 8, 1, 11, 6

$$\bar{x} = A + \frac{\sum d_i}{n} = 37 + \frac{0 + 6 + 5 + \ldots + 11 + 6}{15} = 37 + 4.13 = 41.13$$

For grouped data: $\bar{x} = A + \frac{\sum f_i d_i}{\sum f_i}$

where A is an arbitrary class mark.

Let A = 38

$$\bar{x} = A + \frac{\sum f_i d_i}{\sum f_i} = 38 + \frac{48}{15} = 38 + 3.2 = 41.2$$

Weighted Mean:

Sometimes we associate with the numbers $x_1, x_2, \ldots, x_n$ certain weights $w_1, w_2, \ldots, w_n$ depending on the significance or importance attached to the numbers. In this case,

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n}$$

Example:

Suppose that in a semester you received a grade of 70 in a 3-hour course, 84 in a 4-hour course, and 90 in a 5-hour course. Compute your weighted average semester's grade.

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3}{w_1 + w_2 + w_3} = \frac{3(70) + 4(84) + 5(90)}{3 + 4 + 5} = 83$$

Properties of the Arithmetic Mean:

1. The algebraic sum of deviations of a set of numbers from their arithmetic mean is zero: i.e.;

$$\sum (x_i - \bar{x}) = 0$$

proof :
$$\sum x_i - \sum \bar{x} =$$
$$\sum x_i - n\bar{x} =$$
$$\sum x_i - n\frac{\sum x_i}{n} = 0$$

2. The sum of squares of deviations of a set of numbers x_i from $\bar{x}$ is the minimum: i.e.;

$\sum (x_i - \bar{x})^2 < \sum (x_i - a)^2$, where a is any number $\neq \bar{x}$

Proof:

$$\begin{aligned}\sum(x_i-\bar{x})^2 &= \sum(x_i-a+a-\bar{x})^2 \\ &= \sum\left[(x_i-a)^2+2(x_i-a)(a-\bar{x})+(a-\bar{x})^2\right] \\ &= \sum(x_i-a)^2+2(a-\bar{x})\sum(x_i-a)+n(a-\bar{x})^2 \\ &= \sum(x_i-a)^2+2(a-\bar{x})\left(\sum x_i-na\right)+n(a-\bar{x})^2 \\ &= \sum(x_i-a)^2+2(a-\bar{x})(n\bar{x}-na)+n(a-\bar{x})^2 \\ &= \sum(x_i-a)^2-2n(a-\bar{x})^2+n(a-\bar{x})^2 \\ &= \sum(x_i-a)^2-n(a-\bar{x})^2\end{aligned}$$

$$\therefore \sum(x_i-\bar{x})^2 < \sum(x_i-a)^2$$

Example:

x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	(x_i-5)	$(x_i-5)^2$	$y_i=x_i+5$	$y_i=5x_i$
3	-3	9	-2	4	8	15
5	-1	1	0	0	10	25
9	3	9	4	16	14	45
12	6	36	7	49	17	60
7	1	1	2	4	12	35
4	-2	4	-1	1	9	20
2	-4	16	-3	9	7	10
42		76		83	77	210

3. For any constant a, if $y_i=x_i+a$ *Then* $\bar{y}=\bar{x}+a$

Proof:

$$\bar{y}=\frac{\sum y_i}{n}=\frac{\sum(x_i+a)}{n}=\frac{\sum x_i+na}{n}=\frac{\sum x_i}{n}+a$$

Example:

Let $a = 5$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{42}{7} = 6 \qquad \therefore \bar{y} = \bar{x} + 5 = 6 + 5 = 11$$

4. For any constant a, if $y_i = ax_i$ *Then* $\bar{y} = a\bar{x}$

Proof:

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\sum ax_i}{n} = a\frac{\sum x_i}{n} = a\bar{x}$$

Example:

Let $a = 5$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{42}{7} = 6 \qquad \therefore \bar{y} = 5\bar{x} = 5(6) = 30$$

5. If $z_i = x_i + y_i$ *Then* $\bar{z} = \bar{x} + \bar{y}$

Proof:

$$\bar{z} = \frac{\sum z_i}{n} = \frac{\sum (x_i + y_i)}{n} = \frac{\sum x_i + \sum y_i}{n} = \frac{\sum x_i}{n} + \frac{\sum y_i}{n} = \bar{x} + \bar{y}$$

Example:

Let $z_i = x_i + y_i$

x_i	y_i	z_i
80	60	140
85	63	148
70	50	120
90	74	164
325	247	572

$$\bar{x} = \frac{\sum x_i}{n} = \frac{325}{4} = 81.25 \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{247}{4} = 61.75$$

$$\bar{z} = \frac{\sum z_i}{n} = \frac{572}{4} = 143 \qquad Or \quad \bar{z} = \bar{x} + \bar{y} = 81.25 + 61.75 = 143$$

2. *The Geometric Mean*:

It is defined as the *nth* root of the product of *n* observations.

$$G = \sqrt[n]{x_1 . x_2 . x_3 \ldots x_n}$$

It is of interest in engineering calculations, and is used when dealing with observations each of which bears an approximately constant ratio to the preceding one. For example in averaging rates of growth (increase or decrease) of a statistical population.

In practice, G is computed by logarithms:

$$\log G = \log \left(x_1 . x_2 . x_3 \ldots x_n\right)^{\frac{1}{n}}$$

$$= \frac{1}{n}\left(\log x_1 + \log x_2 + \log x_3 + \ldots + \log x_n\right) = \frac{1}{n}\sum \log x_i$$

Example:

Find the geometric mean of the numbers: 3, 5, 6, 6, 7, 10, and 12

$$G = \sqrt[7]{3.5.6.6.7.10.12} = \sqrt[7]{453600} = 6.43$$

or

$$\log G = \frac{1}{7}\left(\log 3 + \log 5 + \log 6 + \log 6 + \log 7 + \log 10 + \log 12\right) = 0.8081$$

$\therefore G = Anti(\log G) = 6.43$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{3+5+6+6+7+10+12}{7} = 7$$

$G \le \bar{x}$ *always*

Grouped data:

For grouped data the geometric mean is calculated as follows:

$$G = \sqrt[n]{x_1^{f_1} . x_2^{f_2} \ldots x_k^{f_k}} \qquad \text{where} \qquad n = \sum_{i=1}^{k} f_i$$

Or

$$\log G = \frac{1}{n} \log \left(x_1^{f_1} . x_2^{f_2} \ldots x_k^{f_k} \right)$$

$$= \frac{1}{n} \left(f_1 \log x_1 + f_2 \log x_2 + \ldots + f_k \log x_k \right)$$

$$= \frac{1}{n} \sum_{i=1}^{k} f_i \log x_i$$

Where: $x_1, x_2, \ldots, x_k$ are class marks

$f_1, f_2, \ldots, f_k$ are class frequencies

Example:

Find the geometric mean for the following frequency distribution:

Class	f_i	x_i	$\log x_i$	$f_i \log x_i$	$f_i x_i$	d_i	$f_i d_i$
155 – 160	7	157.5	2.1973	15.3811	1102.5	-12	-84
161 – 166	2	163.5	2.2135	4.4270	327.0	- 6	-12
167 – 172	4	169.5	2.2292	8.9167	678.0	0	0
173 – 178	1	175.5	2.2443	2.2443	175.5	6	6
179 – 184	4	181.5	2.2589	9.0355	726.0	12	48
185 – 190	2	187.5	2.2730	4.5460	375.0	18	36
	20			44.5506	3384		-6

$$\log G = \frac{1}{n}\sum_{i=1}^{k} f_i \log x_i = \frac{1}{20}(44.5506) = 2.2275$$

$$G = 168.9$$

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{3384}{20} = 169.2$$

Or let $A = 169.5$ $\quad d_i = x_i - A$

$$\bar{x} = 169.5 + \frac{1}{20}(-6) = 169.2$$

3. ***Median***:

The median of a set of observations is the middle observation when they are ranked or arranged in order of magnitude. That is, the median for ungrouped data is the value of the $\left(\frac{n+1}{2}\right)^{th}$ item in the data array, if n is odd. If n is even the median is taken as one-half the sum of the two "middle" values $\left(\frac{n}{2}, \frac{n}{2}+1\right)$.

Example 1:

Find the median of 10, 12, 20, 17, 11, 16, and 15

10, 11, 12, 15, 16, 17, 20

$$\frac{n+1}{2} = \frac{7+1}{2} = 4$$

Median = 15

Example 2:

Find the median of 8, 6, 9, 5, 5, 12, 10, and 7

5, 5, 6, 7, 8, 9, 10, 12

$\frac{n}{2} = 4, \quad \frac{n}{2} + 1 = 5$

$\therefore Median = \frac{7 + 8}{2} = 7.5$

Grouped data:

For grouped data, the median is given by the formula:

$$Median = L + \left(\frac{n/2 - F_{m-1}}{f_m} \right) c$$

where

L = lower limit of the median class (the class that contains the middle item of the distribution).

n = number of observations $n = \sum f_i$

F_{m-1} = sum of the frequencies up to but not including the median class.

f_m = frequency of the median class.

c = class size.

Geometrically, the median is the value of x that divides the histogram or a frequency polygon or curve into two equal areas.

Because the median is a positional value, it is less affected by extreme values than the mean. This property of the median makes it in some cases a useful measure of central tendency.

Example:

Find the median for the following frequency distribution:

Class	f_i	F_i	x_i	$d_i = x_i - 67.6$	$f_i d_i$
54 – 57	3	3	55.5	-12	-36
58 – 61	5	8	59.5	-8	-40
62 – 65	9	17	63.5	-4	-36
66 – 69	12	29	67.5	0	0
70 – 73	5	34	71.5	4	20
74 – 77	4	38	75.5	8	32
78 – 81	2	40	79.5	12	24
	40				-36

The median class is the first class that has cumulative frequency greater than or equal $n/2$.

The median class is: 66 – 69

$$Median = L + \left(\frac{n/2 - F_{m-1}}{f_m} \right) c = 66 + \left(\frac{40/2 - 17}{12} \right) 4$$

$$= 66 + \frac{3}{12}.4 = 66 + 1 = 67$$

$\bar{x} = 66.6$ $G = 66.3273$ $Mode = 67.2$ *or* 67.5

Example:

The hourly wages of five employees in an office are:

2.52, 3.96, 3.28, 19.20, 3.75

Find: 1. Median 2. Mean

1. Arrange in an array 2.52, 3.28, 3.75, 3.96, 19.20

   Median = \$3.7

2. $Mean = \frac{2.52 + 3.28 + 3.75 + 3.96 + 19.20}{5} = \frac{32.71}{5} = \$6.54$

4. ***Mode***:

Is the value of the observation which occurs most frequently, i.e., it is the most common value. The mode may not exist and if it does exist it may not be unique. A distribution having only one mode is called unimodal.

Example:

The set 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18 has mode 9.

The set 3, 5, 8, 10, 12, 15, 16, 20 has no mode.

The set 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9 has modes 4 and 7 and is called bimodal.

Grouped data:

For grouped data, the mode will be the value of x corresponding to the maximum point on the curve. It may represent the class midpoint of the modal class, or it can be obtained by the formula:

$$Mode = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) c$$

where

L: lower limit of modal class (class corresponding to the highest frequency).

Δ_1: difference between modal class and its predecessor.

Δ_2: difference between modal class and its successor.

c: modal class size.

Example:

Following is the distribution of the amount of time spent in the exercise room of a health club by a sample of 75 patrons.

No. minutes	f_i	x_i	d_i	$f_i d_i$
0 – 14	7	7	-30	-210
15 – 29	19	22	-15	-285
30 – 44	27	37	0	0
45 – 59	13	52	15	195
60 – 74	6	67	30	180
75 – 89	3	82	45	135
	75			15

$$Mode = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) c = 30 + \left(\frac{8}{8+14} \right) 15 = 30 + 5.454 = 35.454$$

Or: $Mode = \frac{30 + 44}{2} = 37$

$\bar{x} = 37 + \frac{15}{25} = 37.2$

Median $= 30 + \left(\frac{37.5 - 26}{27} \right) 15 = 30 + 6.389 = 36.389$

Comparison between Mean, Median and Mode:

For unimodal frequency curves which are moderately skewed, we have the empirical relation:

Mean – Mode = 3(Mean – Median)

Or Mean – Median = 1/3(Mean – Mode)

When the three averages do not coincide, the frequency distribution curve is said to be skewed. It is the degree of asymmetry, and in general its value must fall between -3 and 3.

$$Sk = \frac{Mean - Mode}{S.D} = \frac{3(Mean - Median)}{S.D}$$

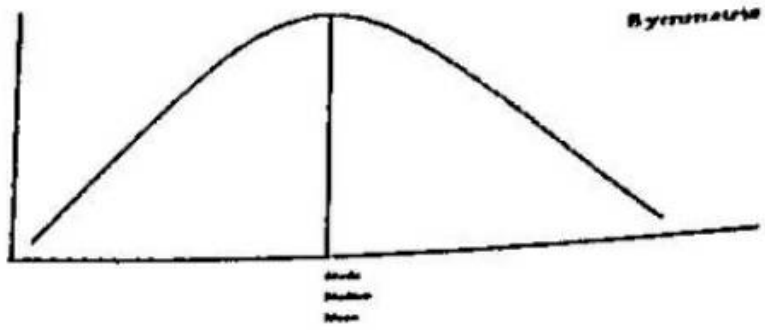
The above two measures are called, respectively, Pearson's first and second coefficients of skewness. It may be used to compare the skewness of different distributions. For symmetrical distributions the value of skewness is zero.

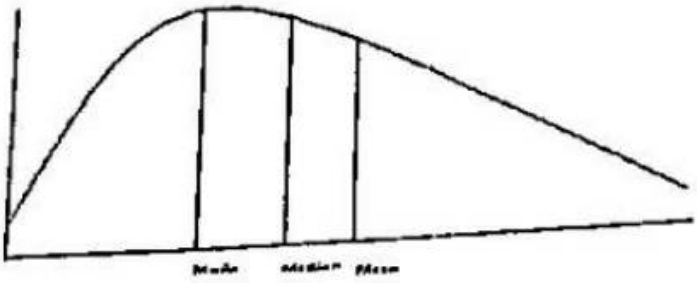
Example:

Find the coefficient of skewness for the distribution which has mean = 56.7, the median = 56.2 and standard deviation = 15.4.

$$Sk = \frac{3(Mean - Median)}{S.D} = \frac{3(56.7 - 56.2)}{15.4} = 0.097$$

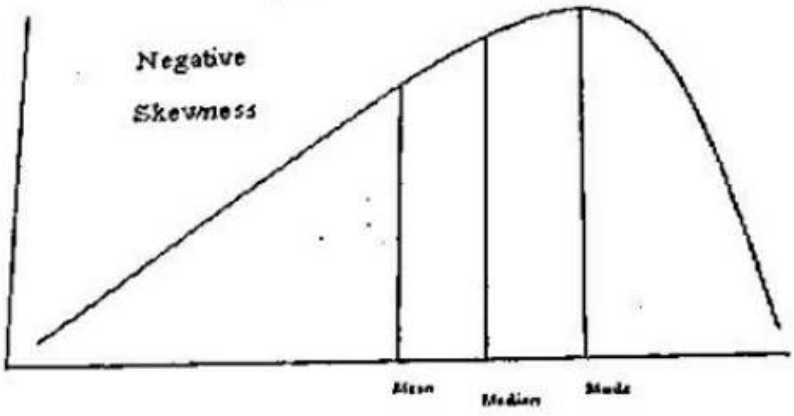
On the basis of this result we can say that the distribution is nearly symmetrical.





Positive

Skewness



Negative

Skewness

Chapter Four

Measures of Variation

The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data. There are several measures of scatter or dispersion, the most common being the following:

1. ***The Range***: It is the difference between the two extremes of the data.

$R = X_n - X_1$

The range is easy to calculate and easy to understand, but it tells us nothing about the dispersion of data which fall between the to extremes. The following sets of data

Set1: 5 17 17 17 17 17 17 17 17 17

Set2 : 5 5 5 5 5 17 17 17 17 17

Set3 :5 6 8 10 11 14 14 15 16 17

has a range of 17 - 5 = 12, but the dispersion is quit different in each case. Nevertheless, the range is a very useful measure of variation when the sample size is quit small. It is used widely in industrial quality control.

2. ***Average Deviation or Mean Deviation***:

It is the mean of the absolute values of devotions:

$$M.D = \frac{\sum |x_i - \bar{x}|}{n}$$

The use of absolute values is necessary because the algebraic sum of deviations from the mean is always zero.

For grouped data:

$$M.D = \frac{\sum f_i |x_i - \bar{x}|}{n}$$

where x_i is the ith class midpoint, $n = \sum f_i$

Example: Find M.D for of numbers 3, 5, 6, 8, 9, 11

$$\bar{x} = \frac{42}{6} = 7$$

$$M.D = \frac{|3-7|+|5-7|+|6-7|+|8-7|+|9-7|+|11-7|}{6} = \frac{14}{6} = 2.333$$

Example: Find *M.D* for the following frequency distribution:

Class	f_i	x_i	$f_i x_i$	$\lvert x_i - \bar{x}\rvert$	$f_i \lvert x_i - \bar{x}\rvert$
37 – 39	7	38	266	3.2	22.4
40 – 42	2	41	82	0.2	0.4
43 – 45	4	44	176	2.8	11.2
46 – 48	2	47	94	5.8	11.6
	15		618		45.6

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{618}{15} = 41.2$$

$$M.D = \frac{\sum f_i |x_i - \bar{x}|}{n} = \frac{45.6}{15} = 3.04$$

3. *Standard Deviation*:

It is the root mean square of the deviation from the mean, and is denoted by s for the sample standard deviation and σ for the population standard deviation.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

For grouped data the standard deviation can be written as:

$$s = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{n}}$$

Sometimes the standard deviation for the data of a sample is defined with $n-1$ replacing n in the denominator; the resulting value represents a better estimate of the standard deviation of a population from which the sample is taken.

For large values of n ($n > 30$) there is practically no difference between the two formulas.

A more convenient form can be obtained using the algebraic identity:

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2$$

$$= \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}\frac{\sum x_i}{n}$$

$$= \sum x_i^2 - \bar{x}\sum x_i = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$s = \sqrt{\frac{1}{n-1}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]}$$ for ungrouped data

And,

$$s = \sqrt{\frac{1}{n-1}\left[\sum f_i x_i^2 - \frac{\left(\sum f_i x_i\right)^2}{n}\right]}$$ for grouped data

The Coding Method:

If $d_i = x_i - A$ are the deviations of x_i from some arbitrary constant A, then

$$s = \sqrt{\frac{1}{n-1}\left[\sum d_i^2 - \frac{\left(\sum d_i\right)^2}{n}\right]}$$ for ungrouped data

And,

$$s = \sqrt{\frac{1}{n-1}\left[\sum f_i d_i^2 - \frac{\left(\sum f_i d_i\right)^2}{n}\right]}$$ for grouped data

Example: Find the standard deviation for the following data:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2	$d_i = x_i - 5$	d_i^2
7	1	1	49	2	4
5	-1	1	25	0	0
8	2	4	64	3	9
4	-2	4	16	-1	1
9	3	9	81	4	16
7	1	1	49	2	4
2	-4	16	4	-3	9
42		36	288	7	43

$$\bar{x} = \frac{\sum x_i}{n} = \frac{42}{7} = 6$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{\frac{36}{7-1}} = \sqrt{6} = 2.45$$

Or,

$$s = \sqrt{\frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]} = \sqrt{\frac{1}{7-1}\left[288 - \frac{(42)^2}{7}\right]} = \sqrt{\frac{36}{6}} = 2.45$$

Or,

$$s = \sqrt{\frac{1}{n-1}\left[\sum d_i^2 - \frac{(\sum d_i)^2}{n}\right]} = \sqrt{\frac{1}{7-1}\left[43 - \frac{(7)^2}{7}\right]} = \sqrt{\frac{36}{6}} = 2.45$$

4. ***Variance***:

Is defined as the square of the standard deviation and is denoted by s^2 for the sample variance and σ^2 for the population variance.

For ungrouped data:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Or, $$s^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]$$

Or, $$s^2 = \frac{1}{n-1}\left[\sum d_i^2 - \frac{(\sum d_i)^2}{n}\right]$$

And, for grouped data:

$$s^2 = \frac{\sum f_i(x_i - \bar{x})^2}{n-1}$$

Or, $$s^2 = \frac{1}{n-1}\left[\sum f_i x_i^2 - \frac{(\sum f_i x_i)^2}{n}\right]$$

Or, $$s^2 = \frac{1}{n-1}\left[\sum f_i d_i^2 - \frac{(\sum f_i d_i)^2}{n}\right]$$

Example1: Find the variance for the following data:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$d_i = x_i - 10$	d_i^2
17	5	25	7	49
3	-9	81	-7	49
10	-2	4	0	0
20	8	64	10	100
10	-2	4	0	0
60		178	10	198

$$\bar{x} = \frac{\sum x_i}{n} = \frac{60}{5} = 12$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{178}{5-1} = 44.5$$

$$\therefore s = \sqrt{44.5} = 6.67$$

Or,

$$s^2 = \frac{1}{n-1}\left[\sum d_i^2 - \frac{\left(\sum d_i\right)^2}{n}\right] = \frac{1}{5-1}\left[198 - \frac{(10)^2}{5}\right] = \frac{178}{4} = 44.5$$

Example2: Find the standard deviation and the variance for the following frequency distribution using the three formulas:

Class	f_i	x_i	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$	x_i^2	$f_i x_i^2$
1 – 5	2	3	6	-9.4	88.36	176.72	9	18
6 – 10	5	8	40	-4.4	19.36	96.8	64	320
11–15	12	13	156	0.6	0.36	4.32	169	2028
16–20	6	18	108	5.6	31.36	188.16	324	1944
	25		310			466		4310

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{310}{25} = 12.4$$

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n-1} = \frac{466}{25-1} = 19.417$$

$$s^2 = \frac{1}{n-1}\left[\sum f_i x_i^2 - \frac{\left(\sum f_i x_i\right)^2}{n}\right] = \frac{1}{25-1}\left[4310 - \frac{(310)^2}{25}\right]$$

$$= \frac{1}{24}[4310 - 3844] = \frac{466}{24} = 19.417$$

Let A = 13

Class	f_i	x_i	$d_i=x_i-13$	f_id_i	$f_id_i^2$
1 – 5	2	3	-10	-20	200
6 – 10	5	8	-5	-25	125
11–15	12	13	0	0	0
16–20	6	18	5	30	150
	25			-15	475

$$s^2 = \frac{1}{n-1}\left[\sum f_i d_i^2 - \frac{\left(\sum f_i d_i\right)^2}{n}\right] = \frac{1}{24}\left[475 - \frac{(-15)^2}{25}\right]$$

$$= \frac{1}{24}[475 - 9] = \frac{466}{24} = 19.417$$

$$\therefore s = \sqrt{19.417} = 4.406$$

H.W: For the following frequency distribution, find: M.D, S.D, and variance using the three formulas.

Class:	28 – 33	34 – 39	40 – 45	46 – 51	52 – 57
f_i:	7	10	9	6	3

Properties of the Standard Deviation and the Variance:

1. If $y_i = x_i + a$, then $s_y^2 = s_x^2$

Proof:

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{\sum [(x_i + a) - (\bar{x} + a)]^2}{n-1} \qquad where \quad \bar{y} = \bar{x} + a$$

$$= \frac{\sum (x_i + a - \bar{x} - a)^2}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1} = s_x^2$$

2. If $yi = ax_i$, then $s_y^2 = a^2 s_x^2$

Proof:

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{\sum (ax_i - a\bar{x})^2}{n-1} \quad where \quad \bar{y} = a\bar{x}$$

$$= \frac{\sum a^2 (x_i - \bar{x})^2}{n-1} = \frac{a^2 \sum (x_i - \bar{x})^2}{n-1} = a^2 s_x^2$$

Example:

x_i	x_i^2	$y_i = x_i + 5$	$z_i = x_i - 2$	$v_i = 4x_i$	$w_i = x_i/2$
-3	9	2	-5	-12	-1.5
6	36	11	4	24	3
5	25	10	8	20	2.5
3	9	8	1	12	1.5
7	49	12	5	28	3.5
18	128				

$$s_x^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right] = \frac{1}{5-1}\left[128 - \frac{(18)^2}{5}\right]$$

$$= \frac{1}{4}[128 - 64.8] = \frac{63.21}{4} = 15.8$$

$\therefore s_y^2 = 15.8$ *and* $s_z^2 = 15.8$,

$s_v^2 = (4)^2(15.8) = 252.8$ *and* $s_w^2 = \frac{1}{4}(15.8) = 3.95$

Coefficient of Dispersion (variation):

It is a measure of relative dispersion, and is denoted by v:

$$v = \frac{S}{\bar{x}} \times 100\ \%$$

It is generally expressed as a percentage.

Note that the coefficient of variation is independent of units. For this reason it is useful in comparing distributions where units may be different. A disadvantage of the coefficient of variation is that it fails to be useful if $\bar{x}$ is close to zero.

Example:

A manufacturer of television tubes has two types of tubes, A and B. The tubes have respective mean lifetimes $\bar{x}_A$ =1495 hours and $\bar{x}_B$ =1875hours and standard deviations S_A=280hours and S_B=310 hours. Which tube has greater relative variation?

$$v_A = \frac{S_A}{\bar{x}_A} \times 100\ \% = \frac{280}{1495} \times 100\ \% = 18.7\%$$

$$v_B = \frac{S_B}{\bar{x}_B} \times 100\ \% = \frac{310}{1875} \times 100\ \% = 16.5\%$$

∴ Tube A has greater relative variation.

Moments

If $x_1, x_2, \ldots, x_n$ are the n values assumed by the variable x, then the quantity

$$m_r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^r}{n}$$

is called the *rth* moment about the arithmetic mean.

Moments can be defined as the arithmetic mean of various powers of deviations taken from the mean of a distribution. These moments are known as central moments. If $r = 1$, $m_1 = 0$. If $r = 2$, $m_2 = s^2$, the variance.

The *rth* moment about any origin A is defined as:

$$m'_r = \frac{\sum_{i=1}^{n}(x_i - A)^r}{n}$$

If $A = 0$, m'_r, is often called the *rth* moment about zero. The first moment about zero with $r = 1$, is the arithmetic mean.

The moments about an origin are known as raw moments.

Example:

1. Find the first four central moments for the following data: 2, 3, 7, 8, and 10.

$$m_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} = 0$$

$$m_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{(2-6)^2 + (3-6)^2 + (7-6)^2 + (8-6)^2 + (10-6)^2}{5}$$

$$= \frac{46}{5} = 9.2$$

$$m_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n} = \frac{(-4)^3 + (-3)^3 + (1)^3 + (2)^3 + (4)^3}{5} = \frac{-18}{5} = -3.6$$

$$m_4 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n} = \frac{256 + 81 + 1 + 16 + 256}{5} = \frac{610}{5} = 122$$

2. Find the first four raw moments about origin zero.

$$m_1' = \frac{\sum_{i=1}^{n}(x_i - 0)}{n} = \frac{2+3+7+8+10}{5} = 6$$

$$m_2' = \frac{\sum_{i=1}^{n}(x_i - 0)^2}{n} = \frac{4+9+49+64+100}{5} = \frac{226}{5} = 45.2$$

$$m_3' = \frac{\sum_{i=1}^{n}(x_i - 0)^3}{n} = \frac{8+27+343+512+1000}{5} = \frac{1890}{5} = 378$$

$$m_4' = \frac{\sum_{i=1}^{n}(x_i - 0)^4}{n} = \frac{2^4+3^4+7^4+8^4+10^4}{5} = \frac{16594}{5} = 3318.8$$

3. Also Find the first four raw moments about origin A = 4.

$$m_1' = \frac{\sum_{i=1}^{n}(x_i - 4)}{n} = \frac{(2-4)+(3-4)+(7-4)+(8-4)+(10-4)}{5} = 2$$

$$m_2' = \frac{\sum_{i=1}^{n}(x_i - 4)^2}{n} = \frac{(-2)^2+(-1)^2+(3)^2+(4)^2+(6)^2}{5} = \frac{66}{5} = 13.2$$

$$m_3' = \frac{\sum_{i=1}^{n}(x_i - 4)^3}{n} = \frac{298}{5} = 59.6$$

$$m_4' = \frac{\sum_{i=1}^{n}(x_i - 4)^4}{n} = \frac{1650}{5} = 330$$

Moments for grouped data

If $x_1, x_2, \ldots, x_k$ occur with frequencies $f_1, f_2, \ldots, f_k$, then:

$$m_r = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^r}{n}$$

$$m_r' = \frac{\sum_{i=1}^{k} f_i(x_i - A)^r}{n}$$

2. From the data given below, find the first four moments about an arbitrary origin, then calculate the first four moments about the mean by applying the above relationships. For $A = 37.5$ $m_1' = 2.61$ $m_2' = 15.57$ $m_3' = 72.43$

Class	30 - 33	33 - 36	36 - 39	39 - 42	42 - 45	45 - 48
Frequency	2	4	26	47	15	6

$m_4' = 655.29$ $m_1 = 0$ $m_2 = s^2 = 87579$

$m_3 =$

$m_4 =$

Skewness:

We study skewness to have an idea about the shape of the curve of a given data. It is the degree of asymmetry of a distribution. For a symmetric distribution Mean = Median = Mode, otherwise it is called a skewed distribution, and such a distribution could either be positively skewed or negatively skewed.

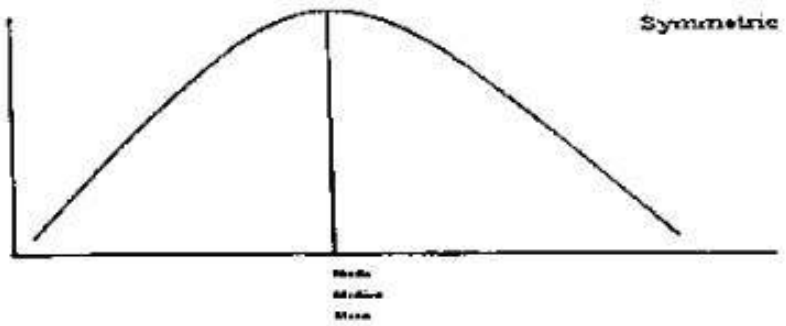
A measure of asymmetry is supplied by the difference (mean – mode). This can be dimensionless on division by a measure of dispersion such as the standard deviation, leading to the following definition:

$$Skewness = \frac{\bar{x} - mode}{s}$$

Or:

$$Skewness = \frac{3(\bar{x} - median)}{s}$$

The above two measures are called, respectively, Pearson's first and second coefficients of skewness.



Relationship between moments:

$m_2 = m_2' - m_1'^2$ where $m_1' = \frac{\sum_{i=1}^{n}(x_i - A)}{n} = \bar{x} - A$

$m_3 = m_3' - 3m_1'm_2' + 2m_1'^3$

$m_4 = m_4' - 4m_1'm_3' + 6m_1'^2m_2' - 3m_1'^4$

In general:

$$m_r = m_r' - C_1^r m_1' m_{r-1}' + C_2^r m_1'^2 m_{r-2}' - \dots + (-1)^{r-1}(r-1)m_1'^r$$

Example:

Prove that: $m_2 = m_2' - m_1'^2$

Let $d_i = x_i - A$ then $x_i = A + d_i$ and $\bar{x} = A + \bar{d}$

$$x_i - \bar{x} = A + d_i - A - \bar{d} = d_i - \bar{d}$$

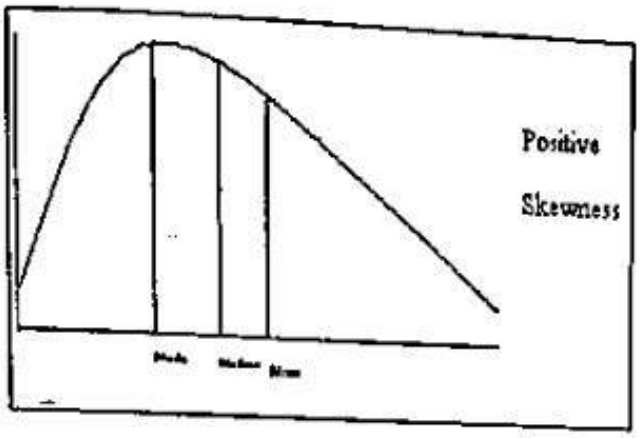
$$m_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n} = \frac{1}{n}\left[\sum d_i^2 - 2\bar{d}\sum d_i + n\bar{d}^2\right]$$

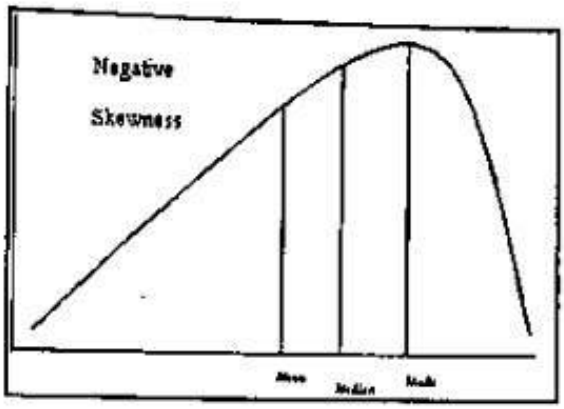
$$= \frac{1}{n}\left[\sum d_i^2 - 2n\bar{d}^2 + n\bar{d}^2\right] = \frac{\sum d_i^2}{n} - \frac{n\bar{d}^2}{n} = \frac{\sum (x_i - A)^2}{n} - (\bar{x} - A)^2$$

$$= m_2' - m_1'^2$$

H.W.:

1. Prove that: $m_3 = m_3' - 3m_1'm_2' + 2m_1'^3$





An important measure of skewness uses the third moment about the mean expressed in dimensionless form is called moment coefficient of skewness.

$$\beta_1 = \frac{m_3}{s^3} = \frac{m_3}{(\sqrt{m_2})^3} = \frac{m_3}{\sqrt{m_2^3}}$$

Kurtosis:

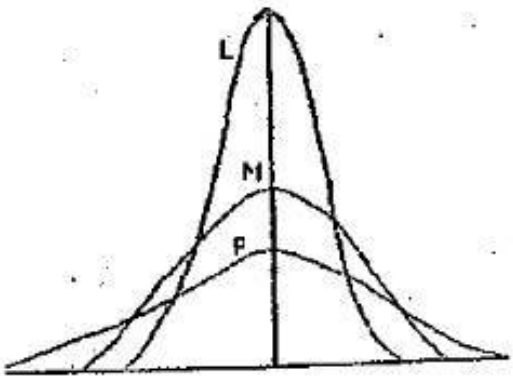
Measure of kurtosis tells as the extent to which a distribution is more peaked or more flat topped than the normal curve, which is symmetrical and bell-shaped. It is the degree of peskiness of a distribution, usually taken relative to a normal distribution. A distribution having relatively high peak is called leptokurtic, while a curve which is flat-topped is called platykurtic. The normal curve is called mesokurtic.

One measure of kurtosis uses forth moment about the mean expressed in dimensionless form is called moment coefficient of kurtosis.

$$\beta_2 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2}$$

For the normal distribution $\beta_2 = 3$. Sometimes kurtosis is measured as the difference $(\beta_2 - 3)$, for this reason the kurtosis for a leptokurtic distribution is

positive ($\beta_2 > 3$ or $\beta_2 - 3 > 0$), and is negative for a platykurtic distribution ($\beta_2 < 3$ or $\beta_2 - 3 < 0$), and zero for the normal distribution.



Example:

Find the moment coefficient of skewness and kurtosis for the following table.

Class mark	5	8	11	14	17	20
Frequency	5	6	6	4	3	2

Solution:

x_i	f_i	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$f_i(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$	$f_i(x_i - \bar{x})^4$
5	5	25	-6	36	180	-216	-1080	1296	6480
8	6	48	-3	9	54	-27	-162	81	486
11	6	66	0	0	0	0	0	0	0
14	4	56	3	9	36	27	108	81	324
17	3	51	6	36	108	216	648	1296	3888
20	2	40	9	81	162	729	1458	6561	13122
Total	26	286			540		972		24300

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{286}{26} = 11$$

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n-1} = \frac{540}{25} = 21.6$$

$$m_3 = \frac{\sum f_i (x_i - \bar{x})^3}{n} = \frac{972}{26} = 37.38$$

$$m_4 = \frac{\sum f_i (x_i - \bar{x})^4}{n} = \frac{24300}{26} = 943.62$$

$$\beta_1 = \frac{m_3}{\sqrt{m_2^3}} = \frac{37.38}{\left(\sqrt{21.6}\right)^3} = \frac{37.38}{100.388} = 0.372$$

$$\beta_2 = \frac{m_4}{m_2^2} = \frac{934.62}{(21.6)^2} = 2.00$$

Example

Find the moment coefficient of skewness and kurtosis for the following frequency distribution.

Class	*60-62*	*63-65*	*66-68*	*69-71*	*72-74*
Frequency	*5*	*18*	*42*	*27*	*8*

$m_2 = s^2 = 8.5275,\ m_3 = -2.6932,\ m_4 = 199.3759$

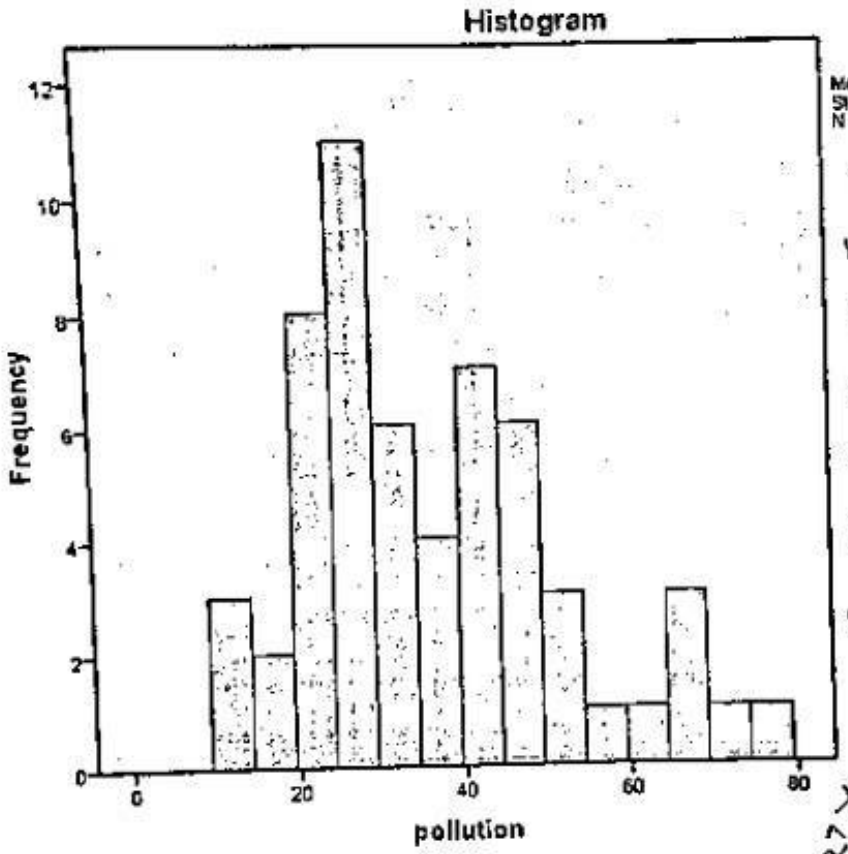
$$\beta_1 = \frac{m_3}{\sqrt{m_2^3}} = \frac{-2.6932}{\left(\sqrt{8.5275}\right)^3} = -0.14$$

$$\beta_2 = \frac{m_4}{m_2^2} = \frac{199.3759}{(8.5275)^2} = 2.74$$

Statistics

pollution

N	Valid	57
	Missing	0
Mean		36.72
Std. Error of Mean		2.101
Median		32.00
Mode		27[a]
Std. Deviation		15.865
Variance		251.706
Skewness		.761
Std. Error of Skewness		.316
Kurtosis		.130
Std. Error of Kurtosis		.623
Range		67
Minimum		12
Maximum		79
Sum		2093



correlation H.W

$r_{12} = .623$

$r_{13} = .359$

$r_{23} = -.126$

$r_{13.2} = -.564$

$r_{12.3} = .722$

$r_{23.1} = -.479$

Regression H.W

$X = \hat{\alpha} + \hat{\beta} Y$

$\hat{\alpha} = 11.168$

$\hat{\beta} = 2.82$

$R^2 = .91$

adjusted $= .9$

ray $= .955$

$1 - r_{13}^2 = 0.87$

$1 - r_{23}^2 = .984$

$1 - r_{12}^2 = .612$

$Y = \hat{\alpha} + \hat{\beta} X$

$\hat{\alpha} = -2.21$ $\hat{\beta} = .3232$

Chapter six

Correlation and Regression

Correlation:

Correlation analysis measures the degree of relationship between the variables.

When only two variables are involved we speak of simple correlation. When more than two variables are involved we speak of multiple correlation.

Linear Correlation:

A first step is the collection of data. Suppose x and y denote respectively the heights and weights of n adult males.

A next step is to plot the points (x_1, y_1), (x_2, y_2), ... , (x_n, y_n). The resulting set of points is called a **scatter diagram.**

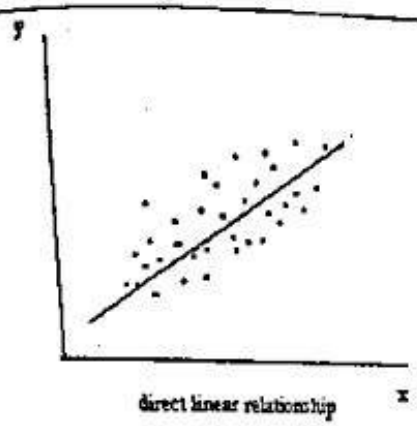
If all points in this scatter diagram seem to lie near a line we say that the correlation is linear.

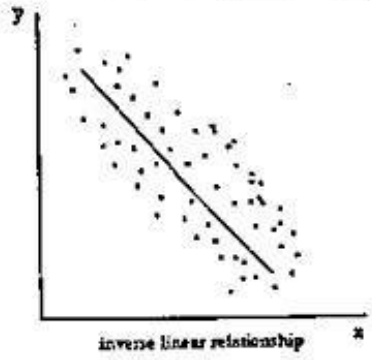
If y tends to increase as x increases the correlation is called positive or direct correlation.

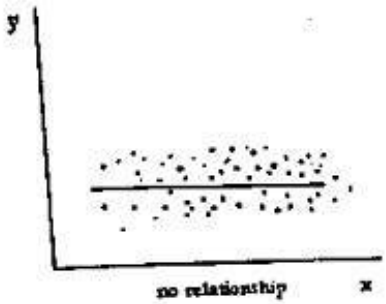
If y tends to decrease as x increases the correlation is called negative or inverse correlation.

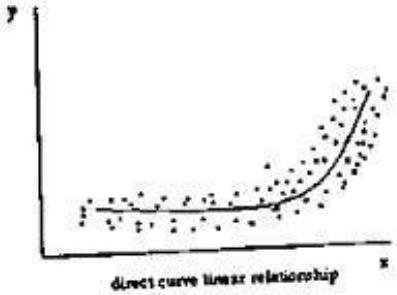
If all points seem to lie near some curve, the correlation is called non-linear.

If there is no relationship indicated between the variables, we say that there is no correlation between them or they are uncorrelated.









Simple correlation coefficient:

If a linear relationship between two variables is assumed, the quantity r called the coefficient of correlation is given by:

$$r = \frac{\text{cov}(x,y)}{s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

Where $cov(x, y) = s_{xy}$ is called the covariance of x and y. The covariance measures the extent to which two variables "vary together". The formula for the sample covariance is:

$$\text{cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \quad , \quad s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}$$

The formula of r becomes:

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

This can be simplified to:

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

Or

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}}$$

The quantity r varies between -1 and +1. The sign ± are used for positive and negative correlation respectively. Note that r is dimensionless quantity.

In order to prove that $|r| \leq 1$, we begin with

$$\sum\left(\frac{x_i - \bar{x}}{s_x} - \frac{y_i - \bar{y}}{s_y}\right)^2 \geq 0$$

$$\sum\frac{(x_i - \bar{x})^2}{s_x^2} + \sum\frac{(y_i - \bar{y})^2}{s_y^2} - 2\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \geq 0$$

Or

$$n - 1 + n - 1 - 2(n-1)\,r \geq 0$$

showing that $r \leq 1$

To show that $r \geq -1$, we start with

$$\sum\left(\frac{x_i - \bar{x}}{s_x} + \frac{y_i - \bar{y}}{s_y}\right)^2 \geq 0$$

And use the same argument given above.

Hence, $-1 \leq r \leq 1$

Example:

Find the coefficient of linear correlation between the variables x and y.

x	y	x^2	y^2	xy	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
1	1	1	1	1	-6	-4	24	36	16
3	2	9	4	6	-4	-3	12	16	9
4	4	16	16	16	-3	-1	3	9	1
6	4	36	16	24	-1	-1	1	1	1
8	5	64	25	40	1	0	0	1	0
9	7	81	49	63	2	2	4	4	4
11	8	121	64	88	4	3	12	16	9
14	9	196	81	126	7	4	28	49	16

$n=8,\ \sum x_i=56,\ \sum y_i=40,\ \sum x_i^2=524,\ \sum y_i^2=256,\ \sum x_i y_i = 364, \sum (x_i-\bar{x})^2=132,\ \sum (y_i-\bar{y})^2=56,\ \sum (x_i-\bar{x})(y_i-\bar{y})=84$

$$r=\frac{\sum[(x_i-\bar{x})(y_i-\bar{y})]}{\sqrt{\sum(x_i-\bar{x})^2(y_i-\bar{y})^2}}=\frac{84}{\sqrt{(132)(56)}}=0.997$$

Or

$$r=\frac{n\sum x_i y_i-\left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left[n\sum x_i^2-\left(\sum x_i\right)^2\right]\left[n\sum y_i^2-\left(\sum y_i\right)^2\right]}}=\frac{(8)(364)-(56)(40)}{\sqrt{\left[(8)(524)-(56)^2\right]\left[(8)(256)-(40)^2\right]}}=0.997$$

Example:

On the basis of the following data, determine whether there is a relationship between the time, in minutes, it takes a secretary to complete a certain form in the morning and in the late afternoon.

Morning(x)	8.2	9.6	7.0	9.4	10.9	7.1	9.0	6.6	8.4	10.5
Afternoon(y)	8.7	9.6	6.9	8.5	11.3	7.6	9.2	6.3	8.4	12.3

$n=10,\quad \sum x_i=86.7,\quad \sum y_i=88.8,\quad \sum x_i^2=771.35,\quad \sum y_i^2=819.34,$
$\sum x_i y_i=792.92$

$$r=\frac{n\sum x_i y_i-\left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left[n\sum x_i^2-\left(\sum x_i\right)^2\right]\left[n\sum y_i^2-\left(\sum y_i\right)^2\right]}}$$

$$=\frac{(10)(792.92)-(86.7)(88.8)}{\sqrt{\left[(10)(771.35)-(86.7)^2\right]\left[(10)(819.34)-(88.8)^2\right]}}=0.936$$

Rank Correlation:

Instead of using precise values of the variables, or when such precision is not available, the data may be ranked in order of size, importance, etc., using the numbers 1, 2, ... , n. If two variables are ranked in such a manner the coefficient of rank correlation also called *Spearman's rank correlation coefficient* is:

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

where

d = differences between ranks of corresponding values of x and y.

n = number of pairs of values (x, y) in the data.

Example:

The following are scores which 12 students obtained in the midterm and final examinations in a course in statistics.

x_i	y_i	R_x	R_y	d_i	d_i^2
71	83	5	10	-5	25
49	62	2	3	-1	1
80	76	7.5	6.5	1	1
73	77	6	8	-2	4
93	89	12	11.5	0.5	0.25
85	74	10	5	5	25
58	48	3	1	2	- 4
82	78	9	9	0	0
64	76	4	6.5	-2.5	6.25
32	51	1	2	-1	1
87	73	11	4	7	49
80	89	7.5	11.5	-4	16

$$r_{rank} = 1-\frac{6\sum d_i^2}{n(n^2-1)} = 1-\frac{6(132.5)}{12(144-1)} = 1-\frac{795}{1716} = 1-0.46 = 0.54$$

Multiple Correlation:

The degree of relationship existing between three or more variables is called multiple correlation. If $R_{1.23}$ is the coefficient of multiple correlation of x_1 on x_2 and x_3. Then

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}}$$

If $R_{2.13}$ is the coefficient of multiple correlation of x_2 on x_1 and x_3. Then

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1-r_{13}^2}}$$

If $R_{3.12}$ is the coefficient of multiple correlation of x_3 on x_1 and x_2. Then

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

A coefficient of multiple correlation lies between 0 and 1. The closer it is to 1 the better is the linear relationship between the variables. The closer it is to 0 the worse is the linear relationship. If the coefficient of multiple correlation is 1, the correlation is called perfect.

Example:

For the data in the following table:

i - Compute r_{12}, r_{13} and r_{23}. ii - Compute $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$.

x_1	64	71	53	67	55	58	77	57	56	51	76	68
x_2	57	59	49	62	51	50	55	48	52	42	61	57
x_3	8	10	6	11	8	7	10	9	10	6	12	9

$\sum x_1 = 753$, $\sum x_2 = 643$, $\sum x_3 = 106$, $\sum x_1^2 = 48139$, $\sum x_2^2 = 34843$, $\sum x_3^2 = 976$, $\sum x_1x_2 = 40830$, $\sum x_1x_3 = 6796$, $\sum x_2x_3 = 5779$

i-

$$r_{12} = \frac{n\sum x_1x_2 - (\sum x_1)(\sum x_2)}{\sqrt{\left[n\sum x_1^2 - (\sum x_1)^2\right]\left[n\sum x_2^2 - (\sum x_2)^2\right]}} = \frac{(12)(40830) - (753)(643)}{\sqrt{\left[(12)(48139) - (753)^2\right]\left[(12)(34843) - (643)^2\right]}}$$

$$= 0.8196$$

$$r_{13} = \frac{n\sum x_1x_3 - (\sum x_1)(\sum x_3)}{\sqrt{\left[n\sum x_1^2 - (\sum x_1)^2\right]\left[n\sum x_3^2 - (\sum x_3)^2\right]}} = \frac{(12)(6796) - (753)(106)}{\sqrt{\left[(12)(48139) - (753)^2\right]\left[(12)(976) - (106)^2\right]}}$$

$$= 0.7689$$

$$r_{23} = \frac{n\sum x_2x_3 - \left(\sum x_2\right)\left(\sum x_3\right)}{\sqrt{\left[n\sum x_2^2 - \left(\sum x_2\right)^2\right]\left[n\sum x_3^2 - \left(\sum x_3\right)^2\right]}} = \frac{(12)(5779) - (643)(106)}{\sqrt{\left[(12)(34843) - (643)^2\right]\left[(12)(976) - (106)^2\right]}}$$

$$= 0.7984$$

ii-

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.8196)^2 + (0.7698)^2 - 2(0.8196)(0.7698)(0.7984)}{1 - (0.7984)^2}}$$

$$= 0.8418$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} = \sqrt{\frac{(0.8196)^2 + (0.7984)^2 - 2(0.8196)(0.7698)(0.7984)}{1 - (0.7698)^2}}$$

$$= 0.8606$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}} = \sqrt{\frac{(0.7698)^2 + (0.7984)^2 - 2(0.8196)(0.7698)(0.7984)}{1 - (0.8196)^2}}$$

$$= 0.8234$$

Note that the coefficient of multiple correlation, such as $R_{1.23}$ is larger than either of the coefficients r_{12} or r_{13}. This is always true, since by taking into account additional variables we should arrive at a better relationship between the variables.

Partial Correlation:

It is often important to measure the correlation between two variables when all other variables involved are kept constant, i.e. when the effects of all other variables are removed. This can be obtained by defining a *coefficient of partial correlation*.

If we denote by $r_{12.3}$ the coefficient of partial correlation between x_1 and x_2 keeping x_3 constant, we find that:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

Example:

Compute the coefficients of linear partial correlation

i) $r_{12.3}$ ii) $r_{13.2}$ iii) $r_{23.1}$ for the data of the previous example.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} = 0.5334$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} = 0.3346$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}} = 0.4580$$

It follows that for constant x_3 the correlation coefficient between x_1 and x_2 is 0.53. For constant x_2 the correlation coefficient between x_1 and x_3 is only 0.33, and finally for constant x_1 the correlation coefficient between x_2 and x_3 is 0.46.

H.W.

The following table reports salary data for n = 14 randomly sampled systems analysts with their years of experience and years of post secondary education.

i. Compute the coefficients of multiple correlation $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$.

ii. Compute the coefficients of partial correlation $r_{12.3}$, $r_{13.2}$, and $r_{23.1}$.

Annual salary$ x_1	years of experience x_2	years of postsecondary education x_3
54900	5.5	4.0
60500	9.0	4.0
58900	4.0	5.0
59000	8.0	4.0
57500	9.5	5.0
55500	3.0	4.0
56000	7.0	3.0
52700	1.5	4.5
56000	8.5	5.0
60000	7.5	6.0
56000	9.5	2.0
53600	6.0	2.0
55000	2.5	4.0
52500	1.5	4.5

Regression

It is frequently desirable to express the relationship between two (or more) variables in mathematical form by determining an equation connecting the variables.

In many situations, there is a single response variable y also called dependent variable, which depends on the value of a set of independent variables $x_1, x_2, \ldots, x_r$. The simplest type of relationship is a linear relationship. That is, for some constants $\beta_0, \beta_1, \ldots, \beta_r$ the equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x_r + e$$

is called a linear regression equation. It describes the regression of y on the set of independent variables $x_1, x_2, \ldots, x_r$. The quantities $\beta_0, \beta_1, \ldots, \beta_r$ are called the regression coefficients and e representing the random error.

A regression equation containing a single independent variable is called a *simple regression equation*, whereas one containing many independent variables is called *multiple regression equation.*

Simple linear regression:

A scatter diagram is a graph in which each plotted point represents an observed pair of values for the independent and the dependent variables. If all the points in the scatter diagram seem to lie near a line we say that a linear relationship exists between the variables.

The simple linear regression model can be expressed as

$$y_i = \alpha + \beta x_i + e_i$$

Where yi: value of the dependent variable in the ith observation.

α: first parameter of the regression equation, which indicates the value of y, when $x = 0$.

β: second parameter of the regression equation, which indicates the slope of the regression line.

x_i: the specified value of the independent variable in the ith observation.

e_i: random sampling error in the ith observation.

The parameters α and β in the regression model are estimated by the values $\hat{\alpha}$ and $\hat{\beta}$ that are based on the sample data.

The linear regression equation based on the sample data is called a regression line of y on x since y is estimated from x.

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

Least Squares Estimators of the Regression Parameters:

To determine estimators of α and β we reason as follows:

If $\hat{\alpha}$ is the estimator of α and $\hat{\beta}$ is the estimator of β then the estimated regression line $\hat{y} = \hat{\alpha} + \hat{\beta}x$ provides the best possible fit to the given data.

Since e_i is the difference between the actual response y_i and the estimated response $\hat{y}_i$, that is: $e_i = y_i - \hat{y}_i$

The least squares estimates of the regression coefficients are the values of $\hat{\alpha}$ and $\hat{\beta}$ for which the quantity

$Q = \sum e_i^2 = \sum \left[y_i - \left(\hat{\alpha} + \hat{\beta} x_i \right) \right]^2$, is minimized

Differentiating partially with respect to $\hat{\alpha}$ and $\hat{\beta}$, and equating these partial derivatives to zero we obtain:

$$\frac{\partial Q}{\partial \hat{\alpha}} = \sum -2 \left[y_i - \left(\hat{\alpha} + \hat{\beta} x_i \right) \right] = 0$$

$$\frac{\partial Q}{\partial \hat{\beta}} = \sum -2 x_i \left[y_i - \left(\hat{\alpha} + \hat{\beta} x_i \right) \right] = 0$$

which yield the following normal equations:

$$\sum y_i = n\hat{\alpha} + \hat{\beta} \sum x_i$$

$$\sum x_i y_i = \hat{\alpha} \sum x_i + \hat{\beta} \sum x_i^2$$

If we let $\bar{y} = \frac{\sum y_i}{n}$ and $\bar{x} = \frac{\sum x_i}{n}$, we can write the first normal equation as

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

Substituting $\hat{\alpha}$ into the second normal equation yields

$$\sum x_i y_i = \left(\bar{y} - \hat{\beta} \bar{x} \right) \sum x_i + \hat{\beta} \sum x_i^2$$

$$= \bar{y} \sum x_i - \hat{\beta} \bar{x} \sum x_i + \hat{\beta} \sum x_i^2$$

Or

$$\hat{\beta} \left(\sum x_i^2 - n\bar{x}^2 \right) = \sum x_i y_i - n\bar{x}\bar{y}$$

So

$$\hat{\beta} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Or

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

Residuals:

The difference between the observed value of y and the fitted value $\hat{y}$ is called the residual for that observation and is denoted by e, that is

$$e_i = y_i - \hat{y}_i$$

The set of residuals for the sample data serve as the basis for calculating the standard error of estimate.

The Standard Error of Estimate:

It is the conditional standard deviation of the dependent variable y given a value of the independent variable x.

$$s_{y/x} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

Note that the numerator is the sum of the squares of the residuals.

An alternative computational formula which does not require determination of each fitted value is

$$s_{y/x} = \sqrt{\frac{\sum y_i^2 - \hat{\alpha} \sum y_i - \hat{\beta} \sum x_i y_i}{n-2}}$$

Example:

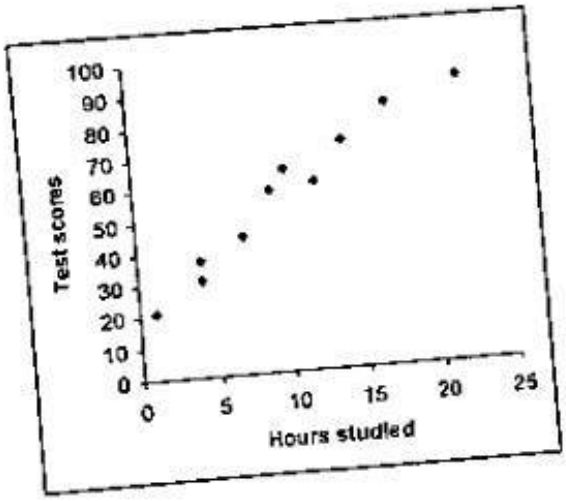
Consider the following 10 data pairs (x_i, y_i), $i = 1, \ldots, 10$, on the number of hours which 10 persons studied for French test and their scores on the test.

i. Find the equation of the least squares line that approximate the regression of the test scores on the number of hours studied.

ii. Predict the average test score of a person who studied 14 hours for the test.

Hours studied(x)	4	9	10	14	4	7	12	22	1	17
Test scores(y)	31	58	65	73	37	44	60	91	21	84

Plotting these data, we get the impression that a straight line provides a reasonably good fit.



i. $n = 10, \sum x_i = 100, \sum x_i^2 = 1376, \sum y_i = 564, \sum x_i y_i = 6945$

$$\hat{\beta} = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2} = \frac{10(6945) - (100)(564)}{10(1376) - (100)^2} = 3.471$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{564}{10} - 3.471\frac{100}{10} = 21.69$$

$\hat{y} = 21.69 + 3.471x$

ii. $\hat{y} = 21.69 + 3.471(14) = 70.284$ Or $\hat{y} = 70$

Note: The arithmetic sign associated with β in the regression equation, indicate the direction of the relationship between x and y (positive = direct, negative= inverse).

H.W.

The following data are measurements of the relative humidity in a storage location and the moisture content of a sample of raw material taken over 15 days.

i. Fit a least squares line and use it to predict the value of moisture content when the relative humidity is 65.

ii. Find $\hat{y}_i$ values corresponding to x_i.

Relative humidity%	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture content%	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

Relationship between Regression and Correlation Coefficients:

The coefficient of the regression line of Y on X or Y given X can be written as:

$$\hat{\beta}_{y|x} = \frac{S_{xy}}{S_x^2}$$

$$S_x \hat{\beta}_{y/x} = \frac{S_{xy}}{S_x} \quad \Rightarrow \quad \frac{S_x}{S_y}\hat{\beta} = \frac{S_{xy}}{S_x S_y} = r_{xy}$$

$$\therefore \hat{\beta}_{y/x} = \frac{S_y}{S_x} r_{xy} \quad Or \quad r_{xy} = \frac{S_x}{S_y}\hat{\beta}_{y/x}$$

Similarly the coefficient of the regression line of X on Y or X given Y can be written as:

$$\hat{\beta}_{x/y} = \frac{S_x}{S_y} r_{xy}$$

The Coefficient of Determination:

It is the square of r_{xy} and is denoted by R^2 used for measuring the proportion of variance in the dependent variable that is statistically explained by the regression equation. R^2 lies between 0 and 1.

The total variation of Y is the sum of the explained variation by the regression equation and the unexplained variation.

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Total sum of Squares = Explained sum of squares + Unexplained sum of squares

$$1 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} + \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \hat{\beta}^2 \frac{S_x^2}{S_y^2}$$

Or,

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

For computational purposes, the following formula is convenient:

$$R^2 = \frac{\hat{\alpha}\sum y_i + \hat{\beta}\sum x_i y_i - n\bar{y}^2}{\sum y_i^2 - n\bar{y}^2}$$

Example:

Given that: $\bar{x} = 3.5$, $\bar{y} = 6$, $r_{xy} = 0.905$, $S_x = 1.333$, $S_y = 2.211$,

find:

1. The regression equation of Y on X

$$\hat{\beta}_{y/x} = \frac{S_y}{S_x} r_{xy} = \frac{2.211}{1.333}(0.905) = 1.501$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 6 - (1.501)3.5 = 6 - 5.2535 \quad = 0.7465$$

$$\therefore \hat{y} = 0.7465 + 1.501x$$

2. The regression equation of X on Y

$$\hat{\beta}_{x/y} = \frac{S_x}{S_y} r_{xy} = \frac{1.333}{2.211}(0.905) = 0.546$$

$$\hat{\alpha} = \bar{x} - \hat{\beta}\bar{y} = 3.5 - (0.546)6 = 3.5 - 3.276 \quad = 0.224$$

$$\therefore \hat{x} = 0.224 + 0.546y$$

3. The coefficient of determination

$$R^2 = (r_{xy})^2 = (0.905)^2 = 0.82$$

Or,

$$R^2 = \hat{\beta}^2{}_{y/x}\frac{S_x^2}{S_y^2} = (1.501)^2\frac{(1.333)^2}{(2.211)^2} = 0.82$$

Or,

$$R^2 = \hat{\beta}^2{}_{x/y}\frac{S_y^2}{S_x^2} = (0.546)^2\frac{(2.211)^2}{(1.333)^2} = 0.82$$