Pimpri Chinchwad Education Trust's

Pimpri Chinchwad College of Engineering

Department of Civil Engineering

ENGINEERING MATHEMATICS III

UNIT III STATISTICS

Mr. Dinesh Kute
(7709495376)

Statistics

Statistics

Def : Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data.

Classification of Data

1. **Ungrouped Frequency Distribution**

2. Grouped Frequency Distribution

Classification of Data

Example : Consider marks of 10 students (out of 25)

23,10,12,23,20,12,20,15,15,25

(Ungrouped Data)

Ungrouped Frequency Distribution

x	f
23	2
10	1
12	2
20	2
15	2
25	1
	$\sum_{i=1}^{6} f_i = N = 10$

Grouped Frequency Distribution

Class	f
0-5	0
5-10	0
10-15	3
15-20	2
20-25	5
	$\sum_{i=1}^{5} f_i = N = 10$

Measures of Central Tendency

Measures of Central tendency :

Measure of central tendency is typical value around which other figures aggregate.

There are five measures of central tendency that are in common use

1) **Arithmetic average or arithmetic mean or simple mean**
2) Median
3) Mode
4) Geometric mean
5) Harmonic mean

Arithmetic mean or Average:

1. Ungrouped Data: Consider set of observations $x_1, x_2, x_3, \ldots \ldots, x_k$

$$\boxed{A.M. = \overline{x} = \frac{\sum_{i=1}^{k} x_i}{k}}$$

Measures of Central Tendency

Arithmetic mean or Average:

2. Ungrouped Frequency distribution: Consider the frequency Distribution

x	f
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_n	f_n

$$A.M = \overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

Measures of Central Tendency

Q.1 Find mean of the following observations

(a) 480,485,492,495,500,505,515

(b)

x	32	33	34	35	36
f	4	10	13	8	5

Solution :

(a) Since, given data is ungrouped data

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i}{k} \quad \Rightarrow \bar{x} = \frac{480+485+492+495+500+505+515}{7} \quad \boxed{\Rightarrow \bar{x} = 496}$$

(b) Since, given data is in ungrouped frequency distribution

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \quad \Rightarrow \bar{x} = \frac{(32\times 4)+(33\times 10)+(34\times 13)+(35\times 8)+(36\times 5)}{4+10+13+8+5} \quad \boxed{\Rightarrow \bar{x} = 34}$$

Arithmetic mean or Average:

3. Grouped Frequency distribution: Consider the frequency Distribution

Class	f	x_i
$a_1 - a_2$	f_1	$\frac{a_1 + a_2}{2}$
$a_2 - a_3$	f_2	$\frac{a_2 + a_3}{2}$
: :	: :	: :
$a_n - a_{n+1}$	f_n	$\frac{a_n + a_{n+1}}{2}$

$$\boxed{A.M. = \bar{x} = A + h\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i} = A + h\frac{\sum_{i=1}^{n} f_i u_i}{N}}$$

Where, $h : lenght\ of\ the\ interval$

$A : middle\ value\ of\ x_i$

$$u_i = \frac{x_i - A}{h}$$

Q.2 The marks obtained in paper of Mathematics are given in the following table. Find the Arithmetic mean of the distribution.

Marks Obtained	No. of students
0-10	8
10-20	20
20-30	14
30-40	16
40-50	20
50-60	25
60-70	13
70-80	10
80-90	5
90-100	2

Solution :

Measures of Central Tendency

Marks Obtained	Mid Value x	No. of students f	$u = \frac{x - A}{h}$	fu
0-10	5	8	-4	-32
10-20	15	20	-3	-60
20-30	25	14	-2	-28
30-40	35	16	-1	-16
40-50	45	20	0	0
50-60	55	25	1	25
60-70	65	13	2	26
70-80	75	10	3	30
80-90	85	5	4	20
90-100	95	2	5	10
Total		$\sum f = 133$		$\sum fu = -25$

$A =$ middle value of $x_i = 45$

$h = 10$

We have

$$\bar{x} = A + h\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i}$$

$$\Rightarrow \bar{x} = 45 + 10\left(\frac{-25}{133}\right)$$

$$\boxed{\Rightarrow \bar{x} = 43.12}$$

Dispersion

Disperson

In some cases average will not give the full information. In this we check how much is scatter/ Disperse from its mean value. To do this we use following Methods of Dispersion

1. **Standard Deviation**

2. **Coefficient of Variation**

Dispersion

Standard deviation

1. Ungrouped Data: Consider set of observations $x_1, x_2, x_3, \ldots \ldots, x_k$

$$Standard\ Deviation = \sigma = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{x})^2}$$

2. Ungrouped Frequency Distribution : Consider the frequency distribution

x	f
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_n	f_n

$$Standard\ Deviation = \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}f_i(x_i - \overline{x})^2}$$

Dispersion

Remarks

1) $Standard\ Deviation = \sqrt{variance}$ $\therefore\ \sigma^2 = variance$

2) For calculation purpose use following formulae for Standard Deviation

- Ungrouped Data

$$\boxed{\sigma^2 = \frac{1}{k}\sum_{i=1}^{k} x_i^2 - (\overline{x})^2} \qquad \sigma = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{x})^2}$$

- Ungrouped Frequency Distribution

$$\boxed{\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - (\overline{x})^2} \qquad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \overline{x})^2}$$

Coefficient of Variation (C.V.) :

$$C.V. = \frac{standard\ Deviation}{A.M.} \times 100 \quad = \frac{\sigma}{\bar{x}} \times 100$$

Remarks

2) For calculation purpose use following formulae for Standard Deviation

- Ungrouped Data

$$\sigma^2 = \frac{1}{k}\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2 \quad = \frac{1}{k}\sum_{i=1}^{k} x_i^{\,2} - 2x_i\bar{x} + \left(\bar{x}\right)^2$$

$$= \frac{1}{k}\sum_{i=1}^{k} x_i^{\,2} - \frac{2}{k}\sum_{i=1}^{k} x_i\bar{x} + \frac{1}{k}\sum_{i=1}^{k}\left(\bar{x}\right)^2$$

$$= \frac{1}{k}\sum_{i=1}^{k} x_i^{\,2} - 2\left(\bar{x}\right)^2 + \left(\bar{x}\right)^2$$

$$= \frac{1}{k}\sum_{i=1}^{k} x_i^{\,2} + \left(\bar{x}\right)^2$$

Dispersion

Remarks

1. Standard Deviation gives information about distribution of data from its mean value.

2. Coefficient of Variation gives information about distribution of data from its mean value in percentage

3. If the standard deviation is 0.20 and the mean is 0.50, then the C.V. = 40%. So CV helps us see that even a lower standard deviation doesn't mean less variable data.

1. More C.V. $\Rightarrow$ More Variability$\Rightarrow$ Less Consistency

   Less C.V. $\Rightarrow$ Less Variability$\Rightarrow$ More Consistency

4. More σ $\Rightarrow$ More Variability$\Rightarrow$ Less Consistency

   Less σ $\Rightarrow$ Less Variability$\Rightarrow$ More Consistency

Examples :

Q.1 Find the standard deviation for the following data

49,63,46,59,65,52,60,54

Solution :

Standard deviation for ungrouped data is $\sigma^2 = \frac{1}{k}\sum_{i=1}^{k} x_i^2 - (\overline{x})^2$ where, $\overline{x} = \frac{\sum_{i=1}^{k} x_i}{k}$

x	x^2
49	2401
63	3969
46	2116
59	3481
65	4225
52	2704
60	3600
54	2916
$\sum = 448$	$\sum = 25412$

$$\therefore \sigma^2 = \frac{1}{8} \times 25412 - \left(\frac{448}{8}\right)^2 = 40.5$$

$$\therefore \sigma = 6.3640$$

Q.2 The scores obtained by two batsman A and B in 10 matches are given below. Determine who is more consistent?

Batsman A	30	44	66	62	60	34	80	46	20	38
Batsman B	34	46	70	38	55	48	60	34	45	30

Solution :

Here, we have to find coefficient of variation of Batsman A i.e. $(C.V.)_A$ & coefficient of variation of Batsman B i.e. $(C.V.)_B$

Where,

$$\text{Coefficient of variation (A)} = (C.V.)_A = \frac{\sigma_A}{(\bar{x})_A} \times 100$$

$$\text{Coefficient of variation (B)} = (C.V.)_B = \frac{\sigma_B}{(\bar{x})_B} \times 100$$

$$\sigma^2 = \frac{1}{k}\sum_{i=1}^{k} x_i^2 - \left(\bar{x}\right)^2 \qquad \bar{x} = \frac{\sum_{i=1}^{k} x_i}{k}$$

Dispersion

Case I : Coefficient of variation (A) $= (C.V.)_A = \frac{\sigma_A}{(\bar{x})_A} \times 100$

Batsman A (x)	x^2
30	900
44	1936
66	4356
62	3844
60	3600
34	1156
80	6400
46	2116
20	400
38	1444
$\sum =$ 480	$\sum =$ 26152

$$\sigma_A^{\ 2} = \frac{1}{k}\sum_{i=1}^{k} x_i^{\ 2} - \left(\bar{x}\right)^2$$

$$\Rightarrow \sigma_A^{\ 2} = \frac{1}{10} \times 26152 - \left(\frac{480}{10}\right)^2$$

$$\Rightarrow \sigma_A^{\ 2} = 311.2$$

$$\Rightarrow \sigma_A = 17.6409$$

$$\boxed{\therefore \sigma_A = 17.6409}$$

$$\boxed{\& \ \left(\bar{x}\right)_A = \frac{480}{10} = 48}$$

$$(C.V.)_A = \frac{\sigma_A}{\left(\bar{x}\right)_A} \times 100$$

$$\therefore (C.V.)_A = \frac{17.6409}{48} \times 100$$

$$\boxed{\therefore (C.V.)_A = 36.7519}$$

Dispersion

Case II : Coefficient of variation (B) $= (C.V.)_B = \frac{\sigma_B}{(\bar{x})_B} \times 100$

Batsman B (x)	x^2
34	1156
46	2116
70	4900
38	1444
55	3025
48	2304
60	3600
34	1156
45	2025
30	900
$\sum =$ 460	$\sum =$ 22626

$$\sigma_B^{\,2} = \frac{1}{k}\sum_{i=1}^{k} y_i^{\,2} - \left(\bar{y}\right)^2$$

$$\Rightarrow \sigma_B^{\,2} = \frac{1}{10} \times 22626 - \left(\frac{460}{10}\right)^2$$

$$\Rightarrow \sigma_B^{\,2} = 146.6$$

$$\Rightarrow \sigma_B = 12.1078$$

$$\boxed{\therefore \sigma_B = 12.1078}$$

$$\boxed{\& \ \left(\bar{x}\right)_B = \frac{460}{10} = 46}$$

$$(C.V.)_B = \frac{\sigma_B}{\left(\bar{x}\right)_B} \times 100$$

$$\therefore (C.V.)_B = \frac{12.1078}{46} \times 100$$

$$\boxed{\therefore (C.V.)_B = 26.3214}$$

Dispersion

Since

$$\boxed{(C.V.)_A = 36.7519 > (C.V.)_B = 26.3214}$$

Therefore, Batsman A has more variability than Batsman B

i.e. Batsman B is more consistent than Batsman A

Dispersion

Q.3 Arithmetic mean and standard deviation of 30 items are 20 and 3 respectively out of these 30 items, item 22 and 15 are dropped. Find new A.M. and S.D. Calculate A.M. and S.D. if item 22 is replaced by 8 and 15 is replaced by 17.

Solution :

Let $x_1, x_2, x_3, \ldots 22, 15, \ldots . x_{30}$ be given items.

Given that,

A.M. $(\bar{x})$ and S.D. (σ) are 20 & 3 respectively

$$\therefore \bar{x} = 20 \;\&\; \sigma = 3 \ldots ..(1)$$

$$i.e.\ \bar{x} = \frac{\sum_{i=1}^{k} x_i}{k} \quad \& \quad \sigma^2 = \left(\frac{1}{k}\sum_{i=1}^{k} x_i^{\,2}\right) - \left(\bar{x}\right)^2$$

$$i.e.\ \frac{\sum_{i=1}^{30} x_i}{30} = 20 \quad \& \quad \frac{1}{30}\sum_{i=1}^{30} x_i^{\,2} - (20)^2 = 9 \qquad \text{From equation (1)}$$

$$i.e.\ \frac{\sum_{i=1}^{28} x_i + 22 + 15}{30} = 20 \quad \& \quad \frac{1}{30}\left[\sum_{i=1}^{28} x_i^{\,2} + 22^2 + 15^2\right] - (20)^2 = 9$$

$$i.e.\ \sum_{i=1}^{28} x_i = 563 \quad \& \quad \sum_{i=1}^{28} x_i^{\,2} = 11561$$

Case I : when item 22 & 15 are dropped

$$\therefore\ \bar{x} = \frac{\sum_{i=1}^{k} x_i}{k}$$

$$\Rightarrow \bar{x} = \frac{\sum_{i=1}^{28} x_i}{28} = \frac{563}{28} = 20.1071 \quad \boxed{\Rightarrow \bar{x} = 20.1071}$$

Find new A.M. and S.D.
Calculate A.M. and S.D. if item 22 is replaced by 8 and 15 is replaced by 17.

Dispersion

$$\sigma^2 = \left(\frac{1}{k}\sum_{i=1}^{k} x_i^{\,2}\right) - \left(\bar{x}\right)^2 \qquad i.e.\ \sum_{i=1}^{28} x_i = 563 \ \ \& \ \ \sum_{i=1}^{28} x_i^{\,2} = 11561$$

$$\Rightarrow \sigma^2 = \frac{1}{28}\left[\sum_{i=1}^{28} x_i^{\,2}\right] - (20.1071)^2$$

$$\Rightarrow \sigma^2 = \frac{1}{28}[11561] - (20.1071)^2$$

$$\Rightarrow \sigma^2 = 8.5974$$

$$\Rightarrow \sigma = 2.9321$$

Case II : when item 22 is replaced by 8 & 15 is replaced by 17

Now, consider the items

$$x_1, x_2, x_3, \ldots 8, 17, \ldots . x_{30}$$

Dispersion

$$\therefore \; \bar{x} = \frac{\sum_{i=1}^{k} x_i}{k}$$

$$i.e. \; \sum_{i=1}^{28} x_i = 563 \;\; \& \;\; \sum_{i=1}^{28} x_i^{\,2} = 11561$$

$$\Rightarrow \; \bar{x} = \frac{\sum_{i=1}^{28} x_i + 8 + 17}{30} = \frac{563+8+17}{30} = 19.6 \quad \boxed{\Rightarrow \; \bar{x} = 19.6}$$

$$and \;\; \sigma^2 = \left(\frac{1}{k} \sum_{i=1}^{k} x_i^{\,2} \right) - \left(\bar{x} \right)^2$$

$$\sigma^2 = \frac{1}{30}\left[\sum_{i=1}^{28} x_i^{\,2} + 8^2 + 17^2 \right] - (19.6)^2$$

$$\sigma^2 = \frac{1}{30}\left[11561 + 8^2 + 17^2 \right] - (19.6)^2$$

$$\sigma^2 = 12.9733 \quad \boxed{\Rightarrow \sigma = 3.6019}$$

Moment :

The r^{th} moment about mean $(\bar{x})$ is denoted by μ_r and defined as

$$\mu_r = \frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{x})^r$$

Also, the r^{th} moment about any Number 'A' is denoted by μ'_r and defined as

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

Relation between μ_r & μ'_r

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2\mu_2' - 3(\mu_1')^4$$

Moments, Skewness and kurtosis

Remarks :

(1) Mean($\bar{x}$)

$$mean = \bar{x} = A + \mu_1{}'$$

(2) Standard Deviation(σ)

$$\text{standard deviation}\left(\sigma\right) = \sqrt{\mu_2}$$

(3) Coefficient of skewness(β_1)

$$\beta_1 = \frac{\mu_3{}^2}{\mu_2{}^3}$$

(4) Coefficient of kurtosis(β_2)

$$\beta_2 = \frac{\mu_4}{\mu_2{}^2}$$

Moments, Skewness and kurtosis

Type	Description	Example	Result
Arithmetic mean	*Sum of values of a data set divided by number of values*	(1+2+2+3+4+7+9) / 7	**4**
Median	*Middle value separating the greater and lesser halves of a data set*	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	*Most frequent value in a data set*	1, **2**, **2**, 3, 4, 7, 9	**2**

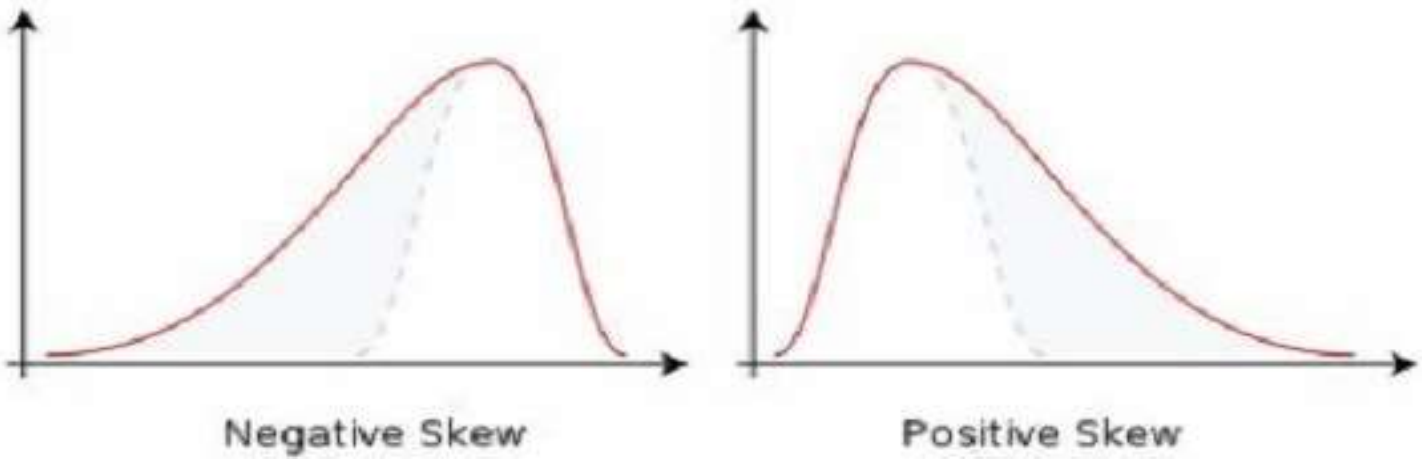
Moments, Skewness and kurtosis

Skewness : It gives imbalance and asymmetry from mean of a data distribution

There are two types of skewness

1. Positive Skewness : If the frequency curve of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or to have positive skewed. In a positive skewed distribution, Mean > Median > Mode

2. Negative Skewness : If the frequency curve has a longer tail to the left of the central maximum than to the right, the distribution is said to be skewed to the left or to have negative skewed. In
a negatively skewed distribution, Mode > Median > Mean.



Remarks :

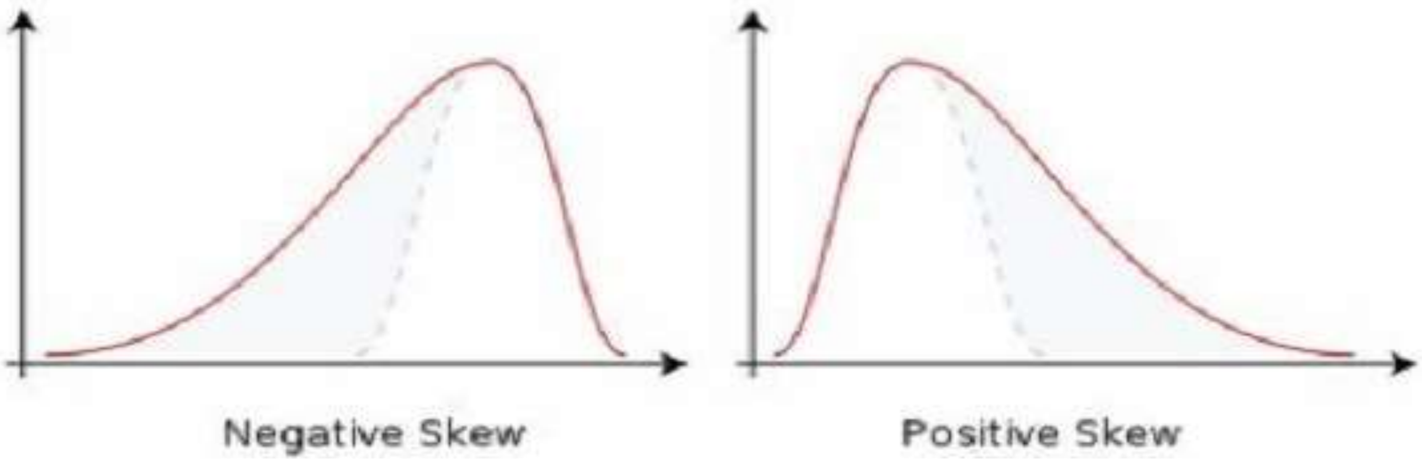
Coefficient of skewness(β_1)

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^2}$$

1. $If\ \beta_1 > 0\ then\ distribution\ is\ called\ Positive\ Skew\ Distribution$

   $If\ \beta_1 < 0\ then\ distribution\ is\ called\ Negative\ Skew\ Distribution$

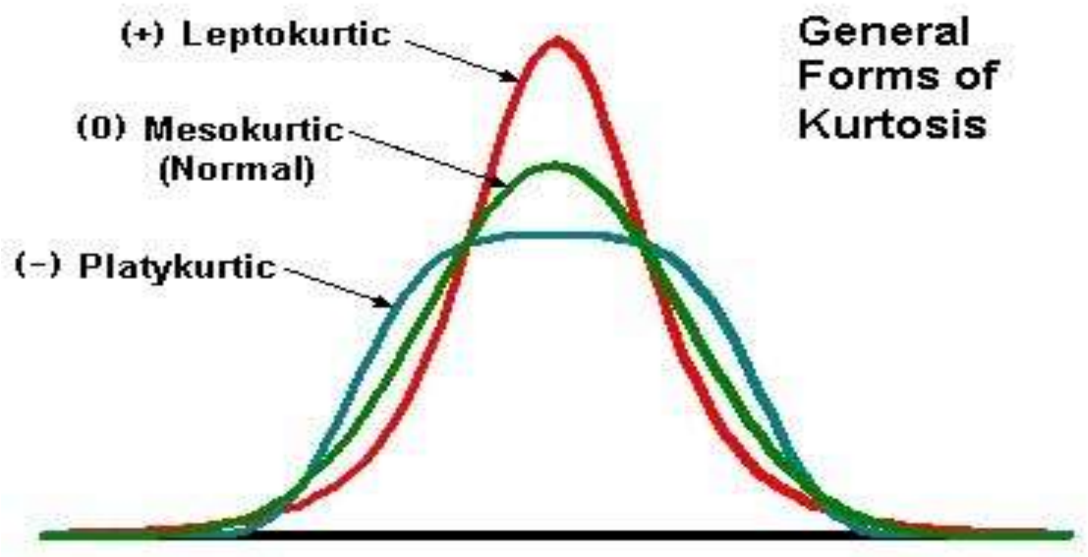
   $If\ \beta_1 = 0\ then\ distribution\ is\ called\ Normal\ Distribution$



Moments, Skewness and kurtosis

Kurtosis : Kurtosis is a measure of peakedness of a distribution relative to the normal distribution(or symmetrical distribution)

A distribution having a relatively high peak is called leptokurtic. A distribution which is flat topped is called platykurtic. The normal distribution which is neither very peaked nor very flat-topped is also called mesokurtic.



Moments, Skewness and kurtosis

Remarks :

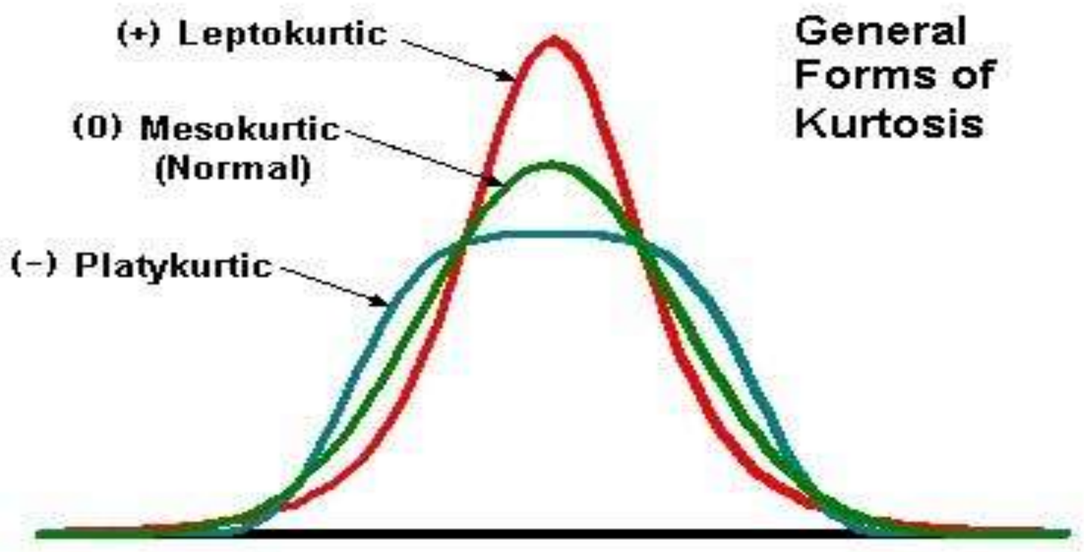
Coefficient of kurtosis(β_2)

$$\beta_2 = \frac{\mu_4}{{\mu_2}^2}$$

2. $If\ \beta_2 > 3\ \ then\ curve\ is\ called\ Leptokurtic$

$If\ \beta_2 = 3\ \ then\ curve\ is\ called\ Mesokurtic$

$If\ \beta_2 < 3\ \ then\ curve\ is\ called\ Platykurtic$



Q.1 The first four moments of a distribution about the value 5 are 2,20,40 and 50. From the given information obtain the first four central moments, mean, standard deviation and coefficients of skewness and kurtosis.

Solution :

Given that

$$A = 5, \mu_1' = 2, \mu_2' = 20, \mu_3' = 40, \mu_4' = 50$$

The first four central moments or moments about mean are

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 \quad = 20 - (2)^2 \quad = 16 \quad \boxed{\Rightarrow \mu_2 = 16}$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 = 40 - 3(2)(20) + 2(2)^3 = -64 \quad \boxed{\Rightarrow \mu_3 = -64}$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2 \mu_2' - 3(\mu_1')^4 = 50 - 4(2)(40) + 6(2)^2(20) - 3(2)^4 = 162$$

$$\boxed{\Rightarrow \mu_4 = 162}$$

Moments, Skewness and kurtosis

Mean

$$mean = \bar{x} = A + \mu_1'$$

$$mean = \bar{x} = 5 + 2 \qquad \boxed{\Rightarrow mean = \bar{x} = 7}$$

Standard Deviation

$$\text{standard deviation} = \sqrt{\mu_2}$$

$$\text{standard deviation} = \sqrt{16} \qquad \boxed{\Rightarrow \text{standard deviation} = 4}$$

Coefficient of skewness

$$\beta_1 = \frac{\mu_3^{\,2}}{\mu_2^{\,2}} \qquad \Rightarrow \beta_1 = \frac{(-64)^2}{(16)^2} \qquad \boxed{\Rightarrow \beta_1 = 1}$$

Coefficient of kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} \qquad \Rightarrow \beta_2 = \frac{162}{(16)^2} \qquad \boxed{\Rightarrow \beta_2 = 0.6328}$$

Q.2 Find the four moments about the mean of the following

x	61	64	67	70	73
f	5	18	42	27	8

also calculate β_1 & β_2.

Solution :

The moment about mean $\bar{x}$ is given below $\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^r$

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)$$

$$\mu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^3$$

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^2$$

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^4$$

Moments, Skewness and kurtosis

x	f	fx	$f(x-\bar{x})$	$f(x-\bar{x})^2$	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
61	5	305	-32.25	208.0125	-1341.6806	8653.84
64	18	1152	-62.1	214.245	-739.1453	2550.051
67	42	2814	-18.9	8.505	-3.8273	1.722263
70	27	1890	68.85	175.5675	447.6971	1141.628
73	8	584	44.4	246.42	1367.6310	7590.352
$\sum = 335$	$N = \sum = 100$	$\sum = 6745$	$\sum = 0$	$\sum = 852.75$	$\sum = -269.3250$	$\sum = 19937.5931$

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=}^{n} f_i} = \frac{6745}{100} = 67.45$$

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right) = \frac{1}{100}\times 0 = 0$$

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^2 = \frac{1}{100}\times 852.75 = 8.5275$$

Moments, Skewness and kurtosis

$$\mu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^3 = \frac{1}{100}\times(-269.3250) = -2.6933$$

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^4 = \frac{1}{100}\times(19937.5931) = 199.3759$$

Coefficient of skewness

$$\beta_1 = \frac{\mu_3^{\,2}}{\mu_2^{\,2}} \Rightarrow \beta_1 = \frac{(-2.6933)^2}{(8.5275)^2} \quad \boxed{\Rightarrow \beta_1 = 0.0998}$$

Coefficient of kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} \Rightarrow \beta_2 = \frac{199.3759}{(8.5275)^2} \quad \boxed{\Rightarrow \beta_2 = 2.7418}$$

Correlation

Correlation :

To measure linear relationship between two variables, Karl Pearson developed a formula called correlation coefficient.

Coefficient of Correlation between two variables x and y denoted by $r(x,y)$ and defined as

$$\boxed{r(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}}$$

Where

1. Ungrouped Data

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

$$\boxed{cov(x,y) = (\frac{1}{n}\sum_{i=1}^{n} x_i y_i) - \overline{x}.\overline{y}}$$

$$\boxed{\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2} \quad \boxed{\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\overline{y})^2}$$

2. Ungrouped Frequency Distribution

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \overline{x})(y_i - \overline{y})$$

$$\boxed{cov(x,y) = \frac{1}{N}\sum_{i=1}^{n} f_i x_i y_i - \overline{x}.\overline{y}}$$

$$\boxed{\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - (\overline{x})^2} \quad \boxed{\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{n} f_i y_i^2 - (\overline{y})^2}$$

Correlation

Remarks :

1. $-1 < r < 1$

2. If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller

EXAMPLES

Q.1 Find the coefficient of correlation for the following table

x	10	14	18	22	26	30
y	18	12	24	6	30	36

Solution :

We know, Karl Pearson's coefficient of correlation $r(x, y)$

$$r(x,y) = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}$$

Correlation

$$where,\ \ \text{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_i y_i-\left(\bar{x}.\bar{y}\right)$$

$$\sigma_x{}^2=\frac{1}{n}\sum_{i=1}^{n}x_i{}^2-\left(\bar{x}\right)^2$$

$$\sigma_y{}^2=\frac{1}{n}\sum_{i=1}^{n}y_i{}^2-\left(\bar{y}\right)^2$$

x	y	xy	x^2	y^2
10	18	180	100	324
14	12	168	196	144
18	24	432	324	576
22	6	132	484	36
26	30	780	676	900
30	36	1080	900	1296
$\sum = 120$	$\sum = 126$	$\sum = 2772$	$\sum = 2680$	$\sum = 3276$

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{120}{6}=20 \qquad \boxed{\Rightarrow \bar{x}=20}$$

$$\bar{y}=\frac{\sum_{i=1}^{n}y_i}{n}=\frac{126}{6}=21 \qquad \boxed{\Rightarrow \bar{y}=21}$$

Correlation

$$\mathrm{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_i y_i-\left(\bar{x}.\bar{y}\right)$$

$$\mathrm{cov}(x,y)=\frac{1}{6}\times 2772-\left(20\times 21\right)$$

$$\boxed{\mathrm{cov}(x,y)=42}$$

x	y	xy	x^2	y^2
10	18	180	100	324
14	12	168	196	144
18	24	432	324	576
22	6	132	484	36
26	30	780	676	900
30	36	1080	900	1296
$\sum=120$	$\sum=126$	$\sum=2772$	$\sum=2680$	$\sum=3276$

$$\boxed{\Rightarrow \bar{x}=20} \qquad \boxed{\Rightarrow \bar{y}=21}$$

$$\sigma_x^{\ 2}=\frac{1}{n}\sum_{i=1}^{n}x_i^{\ 2}-\left(\bar{x}\right)^2$$

$$\sigma_x^{\ 2}=\frac{1}{6}\times 2680-\left(20\right)^2$$

$$\sigma_x^{\ 2}=46.6667 \qquad \boxed{\Rightarrow \sigma_x=6.8313}$$

$$\sigma_y^{\ 2}=\frac{1}{n}\sum_{i=1}^{n}y_i^{\ 2}-\left(\bar{y}\right)^2 \qquad \Rightarrow \sigma_y^{\ 2}=\frac{1}{6}\times 3276-\left(21\right)^2 \qquad \Rightarrow \sigma_y^{\ 2}=105 \qquad \boxed{\Rightarrow \sigma_y=10.2470}$$

Correlation

$$r(x,y)=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}$$

$$\boxed{\operatorname{cov}(x,y)=42}$$

$$\boxed{\Rightarrow \sigma_x=6.8313}$$

$$\boxed{\Rightarrow \sigma_y=10.2470}$$

$$r(x,y)=\frac{42}{(6.8313)(10.2470)}$$

$$\boxed{r(x,y)=0.6}$$

EXAMPLES

Q.2 Find the coefficient of correlation for the following table

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

Solution :

We know, Karl Pearson's coefficient of correlation $r(x,y)$

$$\boxed{\Rightarrow r(x,y)=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}}$$

Correlation

$$where,\ \operatorname{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_i y_i-\left(\bar{x}.\bar{y}\right)$$

$$\sigma_x{}^2=\frac{1}{n}\sum_{i=1}^{n}x_i{}^2-\left(\bar{x}\right)^2$$

$$\sigma_y{}^2=\frac{1}{n}\sum_{i=1}^{n}y_i{}^2-\left(\bar{y}\right)^2$$

x	y	xy	x^2	y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\sum = 30$	$\sum = 40$	$\sum = 214$	$\sum = 220$	$\sum = 340$

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{30}{5}=6 \qquad \boxed{\Rightarrow \bar{x}=6}$$

$$\bar{y}=\frac{\sum_{i=1}^{n}y_i}{n}=\frac{40}{5}=8 \qquad \boxed{\Rightarrow \bar{y}=8}$$

Correlation

$$\text{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-(\bar{x}.\bar{y})$$

$$\text{cov}(x,y)=\frac{1}{5}\times 214-(6\times 8)$$

$$\boxed{\text{cov}(x,y)=-5.2}$$

x	**y**	xy	x^2	y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\Sigma=30$	$\Sigma=40$	$\sum=214$	$\sum=220$	$\sum=340$

$$\boxed{\Rightarrow \bar{x}=6} \qquad \boxed{\Rightarrow \bar{y}=8}$$

$$\sigma_x{}^2=\frac{1}{n}\sum_{i=1}^{n}x_i{}^2-(\bar{x})^2$$

$$\sigma_x{}^2=\frac{1}{5}\times 220-(6)^2$$

$$\sigma_x{}^2=8 \qquad \boxed{\Rightarrow \sigma_x=2.8284}$$

$$\sigma_y{}^2=\frac{1}{n}\sum_{i=1}^{n}y_i{}^2-(\bar{y})^2 \quad \Rightarrow \sigma_y{}^2=\frac{1}{5}\times 340-(8)^2 \quad \Rightarrow \sigma_y{}^2=4 \qquad \boxed{\Rightarrow \sigma_y=2}$$

Correlation

$$r(x,y)=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}$$

$$\boxed{\operatorname{cov}(x,y)=-5.2}$$

$$\boxed{\Rightarrow \sigma_x=2.8284}$$

$$\boxed{\Rightarrow \sigma_y=2}$$

$$r(x,y)=\frac{-5.2}{(2.8284)(2)}$$

$$\boxed{r(x,y)=-0.9192}$$

Q.3 Find the coefficient of correlation for the following table

$$n=20,\sum x=40,\sum x_i^{\ 2}=190,\sum y_i=40,\sum y_i^{\ 2}=200,\sum x_i y_i=150,$$

Solution :

We know that

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{40}{20}=2 \qquad \& \qquad \bar{y}=\frac{\sum_{i=1}^{n}y_i}{n}=\frac{40}{20}=2$$

Correlation

Q.3 Find the coefficient of correlation for the following table

$n=20, \sum x = 40, \sum x_i^2 = 190, \sum y_i = 40, \sum y_i^2 = 200, \sum x_i y_i = 150,$

$\boxed{\bar{x}=2}$ $\boxed{\bar{y}=2}$

$$r(x,y)=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}$$

$$where,\ \operatorname{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\left(\bar{x}.\bar{y}\right)$$

$$=\frac{1}{20}\times 150-(2\times 2)$$

$$=3.5$$

$$\boxed{\Rightarrow \operatorname{cov}(x,y)=3.5}$$

$$\sigma_x^{\ 2}=\frac{1}{n}\sum_{i=1}^{n}x_i^{\ 2}-\left(\bar{x}\right)^2=\frac{1}{20}\times 190-(2)^2=5.5 \qquad \boxed{\Rightarrow \sigma_x=2.3452}$$

$$\sigma_y^{\ 2}=\frac{1}{n}\sum_{i=1}^{n}y_i^{\ 2}-\left(\bar{y}\right)^2=\frac{1}{20}\times 200-(2)^2=6 \qquad \boxed{\Rightarrow \sigma_y=2.4495}$$

$$\Rightarrow r(x,y)=\frac{3.5}{(2.3452\times 2.4495)}$$

$$\boxed{\Rightarrow r(x,y)=0.6093}$$

Regression :

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables.

The line of regression gives best estimate for the value of one variable for some specified value of other variable

Regression line y on x

$$y - \bar{y} = b_{yx}(x - \bar{x}) \qquad \text{where } b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\operatorname{cov}(x,y)}{\sigma_x.\sigma_y}\frac{\sigma_y}{\sigma_x} \qquad \boxed{\Rightarrow b_{yx} = \frac{\operatorname{cov}(x,y)}{\sigma_x^{\,2}}}$$

Regression line x on y

$$x - \bar{x} = b_{xy}(y - \bar{y}) \qquad \text{where } b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\operatorname{cov}(x,y)}{\sigma_x.\sigma_y}\frac{\sigma_x}{\sigma_y} \qquad \boxed{\Rightarrow b_{xy} = \frac{\operatorname{cov}(x,y)}{\sigma_y^{\,2}}}$$

where $\boxed{\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\bar{x}.\bar{y}\right)}$ $\boxed{\sigma_x^{\,2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \left(\bar{x}\right)^2}$ $\boxed{\sigma_y^{\,2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\,2} - \left(\bar{y}\right)^2}$

Remarks :

1) $b_{yx} = r\frac{\sigma_y}{\sigma_x}$, $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ are called regression coefficients

2) $r = \sqrt{b_{yx}.b_{xy}}$

3) If x is given and to find corresponding value of y ,use regression line of y on x

4) If y is given and to find corresponding value of x ,use regression line of x on y

5) If $b_{xy} > 0$ and $b_{yx} > 0$ then $r > 0$

6) If $b_{xy} < 0$ and $b_{yx} < 0$ then $r < 0$

7) Point $(\bar{x}, \bar{y})$ satisfies the equations of lines of regression

$$\boxed{y - \bar{y} = b_{yx}\left(x - \bar{x}\right)} \qquad \boxed{x - \bar{x} = b_{xy}\left(y - \bar{y}\right)}$$

Regression

Q.1 Obtain regression lines for the following data :

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

Solution :

Case I : Regression y on x

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

where

$$b_{yx}=\frac{\operatorname{cov}\left(x,y\right)}{\sigma_x{}^2}$$

Case II : Regression x on y

$$x-\bar{x}=b_{xy}\left(y-\bar{y}\right)$$

where

$$b_{xy}=\frac{\operatorname{cov}\left(x,y\right)}{\sigma_y{}^2}$$

and

$$\boxed{\operatorname{cov}\left(x,y\right)=\frac{1}{n}\sum_{i=1}^{n}x_i y_i-\left(\bar{x}.\bar{y}\right)}$$

$$\boxed{\sigma_x{}^2=\frac{1}{n}\sum_{i=1}^{n}x_i{}^2-\left(\bar{x}\right)^2}$$

$$\boxed{\sigma_y{}^2=\frac{1}{n}\sum_{i=1}^{n}y_i{}^2-\left(\bar{y}\right)^2}$$

Regression

x	y	xy	x^2	y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\sum = 30$	$\sum = 40$	$\sum = 214$	$\sum = 220$	$\sum = 340$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{30}{5} = 6$$

$$\boxed{\Rightarrow \bar{x} = 6}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{40}{5} = 8$$

$$\boxed{\Rightarrow \bar{y} = 8}$$

$$\boxed{\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\bar{x}.\bar{y}\right)}$$

$$\boxed{\sigma_x^{\ 2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\ 2} - \left(\bar{x}\right)^2}$$

$$\boxed{\sigma_y^{\ 2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\ 2} - \left(\bar{y}\right)^2}$$

$$\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\bar{x}.\bar{y}\right) = \frac{1}{5}\times 214 - (6\times 8) = -5.2$$

Regression

$$\Rightarrow \operatorname{cov}(x,y) = -5.2$$

x	y	xy	x^2	y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\sum = 30$	$\sum = 40$	$\sum = 214$	$\sum = 220$	$\sum = 340$

$$\sigma_x^{\ 2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\ 2} - \left(\bar{x}\right)^2 = \frac{1}{5}\times 220 - (6)^2 = 8$$

$$\Rightarrow \sigma_x^{\ 2} = 8$$

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\bar{x}.\bar{y}\right)$$

$$\sigma_y^{\ 2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\ 2} - \left(\bar{y}\right)^2 = \frac{1}{5}\times 340 - (8)^2 = 4$$

$$\sigma_x^{\ 2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\ 2} - \left(\bar{x}\right)^2$$

$$\Rightarrow \sigma_y^{\ 2} = 4$$

$$\sigma_y^{\ 2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\ 2} - \left(\bar{y}\right)^2$$

$$b_{yx} = \frac{\operatorname{cov}(x,y)}{\sigma_x^{\ 2}} = \frac{-5.2}{8} = -0.65 \qquad \Rightarrow b_{yx} = -0.65$$

$$\Rightarrow \bar{x} = 6 \qquad \Rightarrow \bar{y} = 8$$

$$b_{xy} = \frac{\operatorname{cov}(x,y)}{\sigma_y^{\ 2}} = \frac{-5.2}{4} = -1.3 \qquad \Rightarrow b_{xy} = -1.3$$

Regression

Case I : Regression y on x

$$\Rightarrow b_{yx} = -0.65$$

$$\Rightarrow b_{xy} = -1.3$$

$$\Rightarrow \bar{x} = 6 \qquad \Rightarrow \bar{y} = 8$$

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

$$y - 8 = -0.65(x - 6)$$

$$y - 8 = -0.65x + 3.9$$

$$\boxed{y = -0.65x + 11.9}$$

Case II : Regression x on y

$$x - \bar{x} = b_{xy}\left(y - \bar{y}\right)$$

$$x - 6 = -1.3(y - 8)$$

$$x - 6 = -1.3y + 10.4$$

$$\boxed{x = -1.3y + 16.4}$$

Regression

Q.2 Obtain regression lines for the following data :

x	10	14	19	26	30	34	39
y	12	16	18	26	29	35	38

and estimate y for $x = 14.5$ and x for $y = 29.5$

Solution :

Case I : Regression y on x

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

where

$$b_{yx} = \frac{\operatorname{cov}(x, y)}{\sigma_x^{\ 2}}$$

Case II : Regression x on y

$$x - \bar{x} = b_{xy}\left(y - \bar{y}\right)$$

where

$$b_{xy} = \frac{\operatorname{cov}(x, y)}{\sigma_y^{\ 2}}$$

and

$$\boxed{\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\bar{x}.\bar{y}\right)}$$

$$\boxed{\sigma_x^{\ 2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\ 2} - \left(\bar{x}\right)^2}$$

$$\boxed{\sigma_y^{\ 2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\ 2} - \left(\bar{y}\right)^2}$$

Regression

x	y	xy	x^2	y^2
10	12	120	100	144
14	16	224	196	256
19	18	342	361	324
26	26	676	676	676
30	29	870	900	841
34	35	1190	1156	1225
39	38	1482	1521	1444
$\sum = 172$	$\sum = 174$	$\sum = 4904$	$\sum = 4910$	$\sum = 4910$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{172}{7} = 24.5714 \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{174}{7} = 24.8571$$

$$\boxed{\Rightarrow \bar{x} = 24.5714} \qquad \boxed{\Rightarrow \bar{y} = 24.8571}$$

$$\boxed{\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\bar{x}.\bar{y})}$$

$$\boxed{\sigma_x^{\ 2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\ 2} - (\bar{x})^2}$$

$$\boxed{\sigma_y^{\ 2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\ 2} - (\bar{y})^2}$$

$$\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\bar{x}.\bar{y}) = \frac{1}{7} \times 4904 - (24.5714 \times 24.8571)$$

Regression

$$\Rightarrow \operatorname{cov}(x, y) = 89.7977$$

x	y	xy	x^2	y^2
10	12	120	100	144
14	16	224	196	256
19	18	342	361	324
26	26	676	676	676
30	29	870	900	841
34	35	1190	1156	1225
39	38	1482	1521	1444
$\sum = 172$	$\sum = 174$	$\sum = 4904$	$\sum = 4910$	$\sum = 4910$

$$\sigma_x^{\,2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \left(\bar{x}\right)^2 = \frac{1}{7} \times 4910 - \left(24.5714\right)^2$$

$$\Rightarrow \sigma_x^{\,2} = 97.6749$$

$$\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\bar{x}.\bar{y}\right)$$

$$\sigma_y^{\,2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\,2} - \left(\bar{y}\right)^2 = \frac{1}{7} \times 4910 - \left(24.8571\right)^2$$

$$\sigma_x^{\,2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \left(\bar{x}\right)^2$$

$$\Rightarrow \sigma_y^{\,2} = 83.5532$$

$$\sigma_y^{\,2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\,2} - \left(\bar{y}\right)^2$$

$$b_{yx} = \frac{\operatorname{cov}(x, y)}{\sigma_x^{\,2}} = \frac{89.7977}{97.6749} \qquad \Rightarrow b_{yx} = 0.9194$$

$$\Rightarrow \bar{x} = 24.5714$$

$$b_{xy} = \frac{\operatorname{cov}(x, y)}{\sigma_y^{\,2}} = \frac{89.7977}{83.5532} \qquad \Rightarrow b_{xy} = 1.0747$$

$$\Rightarrow \bar{y} = 24.8571$$

Regression

Case I : Regression y on x

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

$$y-24.8571=0.9194\left(x-24.5714\right)$$

$$y-24.8571=0.9194x-22.5909$$

$$\boxed{y=0.9194x+2.2662}$$

For $x = 14.5$

$$y=0.9194\left(14.5\right)+2.2662$$

$$\boxed{y=15.5975}$$

$$\boxed{\Rightarrow b_{yx}=0.9194}$$

$$\boxed{\Rightarrow b_{xy}=1.0747}$$

$$\boxed{\Rightarrow \bar{x}=24.5714}$$

$$\boxed{\Rightarrow \bar{y}=24.8571}$$

and estimate

y for $x=14.5$ and x for $y=29.5$

Regression

Case II : Regression x on y

$$x-\bar{x}=b_{xy}\left(y-\bar{y}\right)$$

$$x-24.5714=1.0747\left(y-24.8571\right)$$

$$x-24.5714=1.0747y-26.7139$$

$$\boxed{x=1.0747y-2.1425}$$

For y = 29.5

$$x=1.0747\left(29.5\right)-2.1425$$

$$\boxed{x=29.5612}$$

$$\boxed{\Rightarrow b_{yx}=0.9194}$$

$$\boxed{\Rightarrow b_{xy}=1.0747}$$

$$\boxed{\Rightarrow \bar{x}=24.5714}$$

$$\boxed{\Rightarrow \bar{y}=24.8571}$$

and estimate

y for $x=14.5$ and x for $y=29.5$

Q.3 The regression equations are $8x-10y+66=0$ and $40x-18y=214$.The value of variance of x is 9. Find

a) The mean values of x and y

b) The correlation x and y and

c) The standard deviation of y

Solution

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

Case I : Mean values of x and y

Since, we know that the point $(\bar{x},\bar{y})$ satisfies both regression lines

$$8\bar{x}-10\bar{y}+66=0 \text{ and } 40\bar{x}-18\bar{y}=214$$

$$\boxed{\bar{x}=13 \text{ and } \bar{y}=17}$$

Regression

Case II : Coefficient of correlation

$$8x-10y+66=0$$

$$40x-18y=214$$

$$r=\sqrt{b_{xy}.b_{yx}}$$

$$\boxed{-1<r<1}$$

Regression y on x : $y-\bar{y}=b_{yx}\left(x-\bar{x}\right)$

Regression x on y : $x-\bar{x}=b_{xy}\left(y-\bar{y}\right)$

To find regression of y on x and regression of x on y

$8x-10y+66=0$	$r=\sqrt{b_{xy}.b_{yx}}$	$40x-18y=214$
$y=\frac{8}{10}x+\frac{66}{10}\quad\Rightarrow b_{yx}=\frac{8}{10}=0.8$	1) $r=\sqrt{0.8\times0.45}$ $\boxed{r=0.6}$	$\cancel{y=\frac{40}{18}x-\frac{214}{18}\Rightarrow b_{yx}=\frac{40}{18}=2.2222}$
$\cancel{x=\frac{10}{8}y-\frac{66}{8}\Rightarrow b_{xy}=\frac{10}{8}=1.25}$	$\cancel{2)\ r=\sqrt{1.25\times2.2222}}$ $\cancel{r=1.6667}$	$x=\frac{18}{40}y+\frac{214}{40}\Rightarrow b_{xy}=\frac{18}{40}=0.45$

Regression

$8x-10y+66=0$	$r=\sqrt{b_{xy}\cdot b_{yx}}$	$40x-18y=214$
$y=\frac{8}{10}x+\frac{66}{10} \Rightarrow b_{yx}=\frac{8}{10}=0.8$	1) $r=\sqrt{0.8\times 0.45}$ $\boxed{r=0.6}$	
		$x=\frac{18}{40}y+\frac{214}{40} \Rightarrow b_{xy}=\frac{18}{40}=0.45$

Regression y on x :

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

$$\Rightarrow b_{yx}=\frac{8}{10}=0.8$$

Regression x on y :

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

$$\Rightarrow b_{xy}=\frac{18}{40}=0.45$$

$$r=\sqrt{0.8\times 0.45} \qquad \boxed{\Rightarrow r=0.6}$$

Case II : standard deviation of y : (σ_x)

$$\text{standard deviation} = \sqrt{\text{variance}} \qquad \Rightarrow \sigma = \sqrt{\text{var}}$$

Given that $var(x) = 9$

$$\Rightarrow \sigma_x = \sqrt{\text{var}(x)}$$

$$\Rightarrow \sigma_x = \sqrt{9} \qquad \boxed{\Rightarrow \sigma_x = 3}$$

Regression

Q.4 If the two lines of regression are $9x+y-\lambda=0$ and $4x+y=\mu$ and the means of x & y are 2 & -3 respectively, find the values of λ,μ and the coefficient of correlation between x & y.

Solution

The regression lines are $9x+y=\lambda$ and $4x+y=\mu$

Since, we know that the point $(\bar{x},\bar{y})$ satisfies both regression lines

$\therefore 9\bar{x}+\bar{y}=\lambda$ and $4\bar{x}+\bar{y}=\mu$

Given that : $\bar{x}=2$ and $\bar{y}=-3$

$\therefore 9(2)+(-3)=\lambda$ and $4(2)+(-3)=\mu$

$\boxed{\lambda=15 \text{ and } \mu=5}$

$\therefore$ The regression lines are $9x+y=15$ and $4x+y=5$

Regression

Case II : Coefficient of correlation

$\boxed{\lambda = 15 \text{ and } \mu = 5}$

$$r = \sqrt{b_{xy}.b_{yx}}$$

Regression y on x : $y - \bar{y} = b_{yx}\left(x - \bar{x}\right)$

$\boxed{-1 < r < 1}$

Regression x on y : $x - \bar{x} = b_{xy}\left(y - \bar{y}\right)$

To find regression of y on x and regression of x on y

$9x + y = 15$	$r = \sqrt{b_{xy}.b_{yx}}$	$4x + y = 5$
$\cancel{y = -9x + 15 \Rightarrow b_{yx} = -9}$	$\cancel{1)\ r = \sqrt{(-9)\times(-0.25)}}$ $\cancel{r = 1.5}$	$y = -4x + 5 \Rightarrow b_{yx} = -4$
$x = \frac{-1}{9}y + \frac{15}{9} \Rightarrow b_{xy} = \frac{-1}{9} = -0.1111$	2) $r = \sqrt{(-0.1111)\times(-4)}$ $\boxed{r = 0.6666}$	$\cancel{x = \frac{-1}{4}y + \frac{5}{4} \Rightarrow b_{xy} = \frac{-1}{4} = -0.25}$

Regression

$9x+y=15$	$r=\sqrt{b_{xy}b_{yx}}$	$4x+y=5$
		$y=-4x+5 \Rightarrow b_{yx}=-4$
$x=\frac{-1}{9}y+\frac{15}{9} \Rightarrow b_{xy}=\frac{-1}{9}=-0.1111$	2) $r=\sqrt{(-0.1111)\times(-4)}$ $\boxed{r=0.6666}$	

Regression y on x :

$$x=\frac{-1}{9}y+\frac{15}{9}$$

$$\Rightarrow b_{xy}=\frac{-1}{9}=-0.1111$$

Regression x on y :

$$y=-4x+5$$

$$\Rightarrow b_{yx}=-4$$

$$r=\sqrt{(-0.1111)\times(-4)}$$

$$r=0.6666$$

Since both regression coefficients b_{xy} and b_{yx} are negative

$\therefore$ we take

$$\boxed{r=-0.6666}$$

Curve Fitting

As a result of certain experiment suppose the values of are (x_i, y_i) for $i = 1,2, \dots n$

Curve fitting is actually the process of constructing curve which has best fit to the above data. Usually these curves are straight line, parabola of second degree & parabola of third degree and so on.

1. Fitting by straight line : $y = ax + b$

Formulae : $$\sum y = a\sum x + nb$$

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

2. Fitting by second degree parabola : $y = ax^2 + bx + c$

Formulae : $$\sum y = a\sum x^2 + b\sum x + nc$$

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

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

Conversion into straight line

$$y = ax^b$$

Taking log_{10} on both side

$$\therefore \log_{10} y = \log_{10} a + \log_{10} x^b$$

$$\therefore \log_{10} y = \log_{10} a + b\log_{10} x$$

Now, let $Y = \log_{10} y, X = \log_{10} x, c = \log_{10} a$

$$\therefore Y = bX + c$$

By curve fitting of straight line

Formulae : $\sum Y = b\sum X + nc$

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

Calculate values of b & c, where $\mathbf{a = 10^c}$

Conversion into straight line

$y = ab^x$

Taking log_{10} on both side

$\therefore \log_{10} y = \log_{10} a + \log_{10} b^x$

$\therefore \log_{10} y = \log_{10} a + x \log_{10} b$

Now, let $Y = \log_{10} y, B = \log_{10} \text{b}, c = \log_{10} a$

$\therefore Y = Bx + c$

By curve fitting of straight line

Formulae : $\sum Y = B \sum x + nc$

$$\sum xY = B \sum x^2 + c \sum x$$

Calculate values of B & c, where $\mathbf{a} = 10^c, b = 10^B$