B.Tech II year

Engineering Mathematics-IV (KAS-402)

Introduction: Measures of central tendency, Moments, Moment generating function (MGF) , Skewness, Kurtosis, Curve Fitting , Method of least squares, Fitting of straight lines, Fitting of second degree parabola, Exponential curves ,Correlation and Rank correlation, Regression Analysis: Regression lines of y on x and x on y, regression coefficients, properties of regressions coefficients and non linear regression.

Module-3 (Statistical Technique-I)

Contents

S. No.	Topic	Page no.
3.1	Introduction : central tendency	2
3.2	The method of moments	6
3.2.1	Skewness and Kurtosis	9
3.2.2	Mathematical expectation	11
3.2.3	Moment generating function	13
3.3	Method of least square	15
3.4	Correlation	18
3.5	Regression	22
3.6	Multiple linear regression	25
	Links for e-contents	27

Applications

- Central tendency, moments, skewness and kurtosis provides an idea to the structure of the distribution
- Correlation and regression is applied to validate and further research, make sound business decisions and drive public initiatives.
- The technique of curve fitting help to analyze the data for prediction purpose.

3.1 Introduction: Central tendency

There are certain features which give a general idea of the distribution and can be determined arithmetically. These are the measures of central tendency for example Arithmetic mean, Median, Mode, Geometric mean, Harmonic mean, Quartile, Percentile and Decile.

Arithmetic mean

If the variable x takes n values $x_1, x_2, x_3, \ldots x_n$ with corresponding frequencies $f_1, f_2, f_3, \ldots f_n$. Then Arithmetic mean is given by

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots x_n f_n}{f_1 + f_2 + \ldots f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} \qquad (1)$$

Example 1: The following frequency distribution gives the cost of production(x) of sugarcane in holdings (f). Obtain the arithmetic mean.

Production range	Cost production (x)	No. of holdings(f)	x.f
2-6	4	1	4
6-10	8	9	72
10-14	12	12	144
14-18	16	20	320
18-22	20	17	340
22-26	24	8	192
26-30	28	11	308

Here, $\sum f_i = 112$ and $\sum f_i x_i = 1380$

Therefore, arithmetic mean $\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1380}{112} = 12.32$

Median

Median is the value of the middle variable when the variable are arranged in increasing or decreasing order. In case of grouped frequency distribution, the median is given by the following simple interpolation:

$$M_d = l + \frac{\frac{N}{2} - c}{f} \times i \qquad (2)$$

where, N is the total number observations

l is the lower limit of the median class

c is the just preceding cumulative frequency

f is the frequency of the median class

i is the width of the class interval

Mode

Mode is the value of the variable which has the maximum frequency. For grouped distribution the mode is given by the following formula

$$M = l + \frac{f - f_{-1}}{2f - f_1 - f_{-1}} \times i \qquad (3)$$

where, N is the total number observations

l is the lower limit of the modal class

f_{-1} is the frequency of the preceding modal class

f_1 is the frequency of the next modal class

f is the frequency of the mode class

i is the width of the class interval

Example 2: Find the median and mode for the following distribution:

Class interval	Frequency	Cumulative frequency
0-5	2	2
5-10	5	7
10-15	7	14
15-20	13	27
20-25	21	48
25-30	16	64
30-35	8	72

For median, $N = 72, \frac{N}{2} = 36$ this lies in 5th class i.e 20-25.

$l = 20, i = 5, c = 27, f = 21$ therefore, from (2), median is

$$M_d = l + \frac{\frac{N}{2} - c}{f} \times i = 20 + \frac{36 - 27}{21} \times 5 = 22.1428$$

Since maximum frequency arises in 5th class for which

$l = 20, f = 21, f_{-1} = 13, f_1 = 16, i = 5$

Now, using (3), mode will be

$$M = l + \frac{f - f_{-1}}{2f - f_1 - f_{-1}} \times i = 20 + \frac{21 - 13}{42 - 13 - 16} \times 5 = 23.08$$

Geometric and Harmonic mean

If the variable x takes n values $x_1, x_2, x_3, \ldots x_n$ with corresponding frequencies $f_1, f_2, f_3, \ldots f_n$.

Then,

Geometric mean is given by

$$G.M = \left(x_1^{f_1}.x_2^{f_2}, \ldots x_n^{f_n}\right)^{1/n} \tag{4}$$

and

Harmonic mean is given byi

$$\frac{1}{H.M} = \frac{1}{n}\sum_{i=1}^{n}\frac{f_i}{x_i} \tag{5}$$

Example 3: The monthly income of 10 families in a certain locality are given below:

A	B	C	D	E	F	G	H	I
85	70	15	75	500	20	45	250	40

Calculate the geometric mean and harmonic mean.

From (4), we have $G.M = \left(x_1^{f_1}.x_2^{f_2}, \ldots x_n^{f_n}\right)^{1/n} \Rightarrow \log G.M = \frac{1}{N}\sum_{i=1}^{n} f_i \log x_i$

Therefore, $\log G.M = \frac{1}{N}\sum_{i=1}^{n} f_i \log x_i = \frac{1}{9}[\log(85.70.15.75.500.20.45.250.40)]$

$$= \frac{1}{9}(1.929 + 1.845 + 1.176 + 1.875 + 2.698 + 1.301 + 1.653 + 2.397 + 1.602)$$

$$= 1.803$$

So that $G.M = 63.59$.

Again, harmonic mean is given by (5)

$$\frac{1}{H.M} = \frac{1}{n}\sum_{i=1}^{n}\frac{f_i}{x_i} = \frac{\frac{1}{85} + \frac{1}{70} + \frac{1}{15} + \frac{1}{75} + \frac{1}{500} + \frac{1}{20} + \frac{1}{45} + \frac{1}{250} + \frac{1}{40}}{9}$$

Which implies $H.M = 42.09,$ on solving.

The partition values: Quartile, Decile and Percentile

For a grouped distribution, consider the following fractile of rth order

$$Z_r = l_r + \frac{Nr - c_r}{f} \times i \qquad (6)$$

where,

l_r is the lower limit of the fractile class

c_r is the total of all the frequencies before the fractile class

f is the frequency of the fractile class

i is the width of the class interval

Now, the above formula gives the median if $r = \frac{1}{2}$, quartiles if $r = \frac{1}{3}, \frac{1}{4}$, deciles if $r = \frac{1}{10}, \frac{2}{10}, \ldots \frac{9}{10}$ and percentiles if $r = \frac{1}{100}, \frac{2}{100}, \ldots \frac{99}{100}$.

Example 4: For the following distribution find the first quartile, 5th decile and 28th percentile:

range	frequency	Cumulative frequency
5-10	2	9
10-15	7	16
15-20	14	30
20-25	18	48
25-30	22	70
30-35	14	84
35-40	16	100

Here, $N = 100, i = 5,$

For first quartile, $r=\frac{1}{4},\frac{N}{4}=25$ which lies in 3rd class interval so $l_r=15,\ f_r=14, c_r=16$

Using (6), $Q_1=l_r+\frac{Nr-c_r}{f}\times i=15+\frac{25-16}{14}\times 5=18.214.$

For 5th decile, $r=\frac{5}{10},\frac{5N}{10}=50$ this lies in 5th class interval so $l_r=25,\ f_r=22, c_r=48$

$$D_5=l_r+\frac{Nr-c_r}{f}\times i=25+\frac{50-48}{22}\times 5=25.454.$$

Also for 28th percentile, $r=\frac{28}{100},\frac{28N}{100}=28$ this lies in 3rd class interval so

$l_r=15,\ f_r=14, c_r=16$

$$P_{28}=l_r+\frac{Nr-c_r}{f}\times i=15+\frac{28-16}{14}\times 5=19.285.$$

Problems for practice

1. From the following table find the mean and median:

x	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45
f	7	10	16	32	24	11	10	5	1

Ans: 20.4 and 20.8 respectively.

2. Calculate the 3rd decile and 20th percentile from the following data:

x	1-6	6-11	11-16	16-21	21-26
f	7	18	25	30	20

Ans: 12 and 9.6 respectively.

3. Calculate the geometric mean of the following distribution:

marks	0-10	10-20	20-30	30-40
f	5	8	3	4

Ans: 14.58.

3.2. The method of moments

Moments are the statistical tools to investigate the nature of the distribution.

If $x_1, x_2, ...x_n$ variables have $f_1, f_2, ...f_3$ frequencies respectively then the rth moment about mean $\bar{x}$ is given by

$$\mu_r = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^r}{N} \quad \text{where} \quad r = 0,1,2,... \quad \text{and} \quad N = \sum f_i \tag{1}$$

Similarly the rth moment about an arbitrary point a is given by

$$\mu_r' = \frac{\sum_{i=1}^{n} f_i (x_i - a)^r}{N} \quad \text{where} \quad r = 0,1,2,... \quad \text{and} \quad N = \sum f_i \tag{2}$$

Again if a =0, we get the moments about origin by the following formula

$$\nu_r = \frac{\sum_{i=1}^{n} f_i x_i^{\,r}}{\sum f_i} \quad \text{where} \quad r = 0,1,2,... \tag{3}$$

Note: μ_1, μ_2 are known as mean and variance respectively, also $\sigma = \sqrt{\mu_2}$ is called standard deviation.

Relation between μ_r' and μ_r

$$\mu_r = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^r}{N} = \frac{1}{N} \sum_{i=0}^{n} f_i [(x_i - a) - (x - a)]^r$$

$$= \frac{1}{N} \sum_{i=0}^{n} f_i [(x_i - a) - \mu_1']^r \quad \text{as} \quad \mu_1' = \bar{x} - a$$

$$= \frac{1}{N} \sum_{i=0}^{n} f_i [(x_i - a)^r - C_1^r (x_i - a)^{r-1} \mu_1'$$

$$+ C_2^r (x_i - a)^{r-2} \mu_1'^2 + ... + (-1)^r \mu_1'^r \,]$$

$$\Rightarrow \mu_r = \mu_r' - C_1^r \mu_{r-1}' + C_2^r \mu_{r-2}' \mu_1'^2 + ... + (-1)^r \mu_1'^r \tag{4}$$

Putting r = 1, 2, 3, 4 ... we get

$$\mu_1 = \frac{\sum f_i x_i}{\sum f_i} - \frac{\bar{x} \sum f_i}{\sum f_i} = 0 \tag{5}$$

$$\mu_2 = \mu_2' - 2\mu_1'^2 + \mu_1'^2 = \mu_2' - \mu_1'^2. \tag{6}$$

$$\mu_3 = \mu_3' - 3\mu_2' \mu_1' + 3\mu_1'^3 - \mu_1'^3 = \mu_3' - 3\mu_2' \mu_1' + 2\mu_1'^3 \tag{7}$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 \tag{8}$$

Similarly we may obtain following relations

$$\mu_2' = \mu_2 + \mu_1'^2. \tag{9}$$

$$\mu_3' = \mu_3 + 3\mu_2'\mu_1' - 2\mu_1'^3 \tag{10}$$

$$\mu_4' = \mu_4 + 4\mu_3'\mu_1' - 6\mu_2'\mu_1'^2 + 3\mu_1'^4 \tag{11}$$

Relation between ν_r and μ_r

We may write ν_r in terms of μ_r as

$$\nu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - a + a)^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i [(x_i - a)^r + C_1^r (x_i - a)^{r-1} a + \ldots + a^r]$$

When $a = \bar{x}$

$$\nu_r = \mu_r + C_1^r \mu_{r-1}\bar{x} + C_2^r \mu_{r-2}\bar{x}^2 + \ldots + \bar{x}^r \tag{12}$$

Putting r = 1, 2, 3, 4,...

$$\nu_1 = \mu_1 + \mu_0\bar{x} = \bar{x} \quad \text{as} \quad \mu_1 = 0, \mu_0 = 1 \text{ obviously} \tag{13}$$

$$\nu_2 = \mu_2 + C_1^2 \mu_1 \bar{x} + C_2^2 \mu_0 \bar{x}^2 = \mu_2 + \bar{x}^2 \tag{14}$$

$$\nu_3 = \mu_3 + C_1^3 \mu_2 \bar{x} + C_2^3 \mu_1 \bar{x}^2 + C_3^3 \mu_0 \bar{x}^3 = \mu_3 + 3\mu_2\bar{x} + \bar{x}^3 \tag{15}$$

$$\nu_4 = \mu_4 + 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 + \bar{x}^4 \tag{16}$$

Example 5: The first four moments of a distribution about x=2 are 1, 2.5, 5.5 and 16. Find the first four moments about the mean and about origin.

Here, $a = 2,\ \mu_1' = 1,\ \mu_2' = 2.5,\ \mu_3' = 5.5,\ \mu_4' = 16$

To find the moments about mean

$$\mu_1 = 0, \quad \mu_2 = \mu_2' - \mu_1'^2 = 1.5,$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 5.5 - 3(2.5) + 2 = 0,$$

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

Now, for the moments about the origin

$$\nu_1 = \bar{x} = a + \mu_1' = 2 + 1 = 3$$

$$\nu_2 = \mu_2 + \bar{x}^2 = 1.5 + (3)^2 = 10.5$$

$$\nu_3 = \mu_3 + 3\mu_2\bar{x} + \bar{x}^3 = 0 + 3(1.5)(3) + (3)^3 = 40.5$$

$$\nu_4 = \mu_4 + 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 + \bar{x}^4 = 168$$

Example 6: Calculate first four moments for the following distribution. Also determine mean and standard deviation

Marks :	5-15	15-25	25-35	35-45	45-55	55-65
Students :	10	20	25	20	15	10

We have following table:

Marks	f	x	fx	$(x-\bar{x})$	$f(x-\bar{x})$	$f(x-\bar{x})^2$	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
5-15	10	10	100	-24	-240	5760	-138240	3317760
15-25	20	20	400	-14	-280	3920	-54880	768320
25-35	25	30	750	-4	-100	400	-1600	6400
35-45	20	40	800	6	120	720	4320	25920
45-55	15	50	750	16	240	3840	61440	983040
55-65	10	60	600	26	260	6760	175760	4569760
Total	100		3400		0	21400	46800	9671200

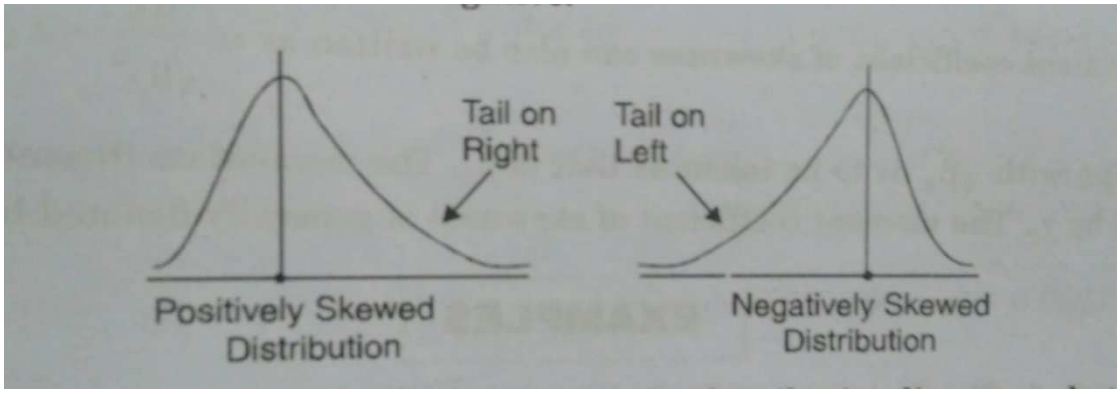
Now, $\mu_1 = \dfrac{\sum f_i(x_i - \bar{x})}{\sum f_i} = 0,$ $\mu_2 = \dfrac{\sum f_i(x_i - \bar{x})^2}{\sum f_i} = \dfrac{21400}{100} = 214,$

$\mu_3 = \dfrac{\sum f_i(x_i - \bar{x})^3}{\sum f_i} = \dfrac{46800}{100} = 468,$ $\mu_4 = \dfrac{\sum f_i(x_i - \bar{x})^4}{\sum f_i} = \dfrac{9671200}{100} = 96712.$

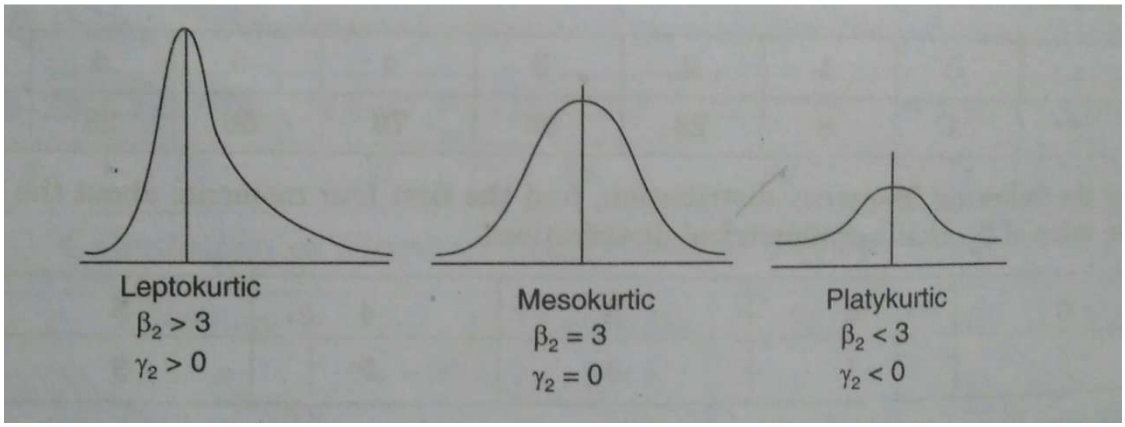
Mean $\bar{x} = \dfrac{\sum f_i x_i}{\sum f_i} = \dfrac{3400}{100} = 34,$ Variance $= \mu_2 = 214 \Rightarrow \sigma = \sqrt{\mu_2} = 14.628$

3.2.1 Skewness and Kurtosis

The lack of symmetry in a distribution is known as skewness.



Kurtosis measures the relative peakedness of the distribution.



Karl Pearson's Coefficients

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \qquad \beta_2 = \frac{\mu_4}{\mu_2^2}, \tag{17}$$

$$\gamma_1 = \sqrt{\beta_1}, \qquad \gamma_2 = \beta_2 - 3. \tag{18}$$

And the moment coefficient of skewness is $\beta_1 = \frac{\mu_3}{\sqrt{\mu_2^3}}$. (19)

Example 7: For a distribution, the mean is 10, variance is 16 and $\beta_1 = 4$. Obtain the first four moments about the origin. Also comment upon the nature of the distribution.

Since, $\mu_1' = \bar{x} = 10, \quad \mu_2 = 16, \quad \gamma_1 = 1, \quad \beta_1 = 4.$

Now $\mu_2' = \mu_2 + \mu_1'^2 = 16 + 100 = 116.$

$$\gamma_1 = 1 \Rightarrow \frac{\mu_3}{\sqrt{\mu_2^3}} = 1 \Rightarrow \mu_3 = 64.$$

$$\mu_3' = \mu_3 + 3\mu_2'\mu_1' - 2\mu_1'^3 = 64 + 3(116)(10) - 2000 = 1544.$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 4 \Rightarrow \mu_4 = 1024.$$

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

$$= 1024 + 4(1544)(10) - 6(116)(100) + 30000 = 23184.$$

Since, $\gamma_1 > 0$, this implies the distribution is positively skewed and $\beta_2 > 3$ implies that the distribution is leptokurtic.

Problems for practice

1. The first four moments of a distribution about the value 4 are 1, 4, 10 and 45. Show that mean is 5 the variance is 3 $\mu_3 = 0$ and $\mu_4 = 26$
2. Compute the first four moments of the data 3, 5, 7, 9 about the mean. Also compute
3. the first four moments about the point 4 Ans: $\mu_1 = 0, \mu_2 = 5, \mu_3 = 0, \mu_4 = 41$ and $\mu_1' = 2, \mu_2' = 9, \mu_3 = 38.25, \mu_4 = 177.$
4. Calculate the first four moments of the following distribution about the mean and hence find β_1, β_2

x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
f	**1**	**8**	**28**	**56**	**70**	**56**	**28**	**8**

Ans: $\mu_1 = 0, \mu_2 = 2, \mu_3 = 0, \mu_4 = 11, \beta_1 = 0, \beta_2 = 2.75$

5. In a frequency distribution the mean is 1.5, variance is 0.64, β_2 =2.5 and $\gamma_1 = 0.3$. Find μ_3, μ_4 and also find first four moments about the origin.

Ans: $\mu_3 = 0.1536, \mu_4 = 1.024, \nu_1 = 1.5, \nu_2 = 2.89, \nu_3 = 6.4086, \nu_4 = 15.6481$

3.2.2 Mathematical expectation

We know that the rth moment about the origin of a distribution with probability function $f(x)$ is given by

$$\mu_r' = \begin{cases} \sum_{i=1}^{n} x_i^r f_i(x_i) \\ \int_{-\infty}^{\infty} x^r f(x) dx \end{cases} \quad \text{for discrete and continuous distribution} \qquad (20)$$

respectively.

And at $r = 1,$ the mathematical expression for computing the expected value of a random variable x with probability function $f(x)$ is given by

$$E(x)=\sum_{i=1}^{n} x_i \; f_i(x_i) \quad \text{or} \quad E(x)=\int_{-\infty}^{\infty} xf(x)dx \tag{21}$$

Also $$E(x^2)=\sum_{i=1}^{n} x_i^{\,2} f_i(x_i) \quad \text{or} \quad E(x^2)=\int_{-\infty}^{\infty} x^2 f(x)dx \tag{22}$$

Example 8: Let x be a random variable with the following data

x	-3	6	9
P(X= x)	1/6	1/2	1/3

Find $E(2x+1)^2$

Since $E(x)=\sum xp(x)=(-3)\frac{1}{6}+\frac{6}{2}+\frac{9}{3}=\frac{11}{2}$ and

$$E(x^2)=\sum x^2 p(x)=9.\frac{1}{6}+\frac{36}{2}+\frac{81}{3}=\frac{93}{2}$$

Now, $$E(2x+1)^2=E(4x^2+4x+1)=4.\frac{93}{2}+4.\frac{11}{2}+1=209.$$

Example 9: Consider the following probability density function

$$f(x)=\begin{cases}\frac{k}{x^3}, x\geq 1\\ 0, else\end{cases}, \text{ find E(x).}$$

First we have to find the value of k

$$\int_1^{\infty}\frac{k}{x^3}dx=1\Rightarrow\frac{k}{2}=1\Rightarrow k=2.$$

$$E(x)=\int_1^{\infty} x\frac{2}{x^3}dx=2$$

Example 10: Find the expectations of numbers obtained on a dice.

We know that to obtain any number has equal probability 1/6 on a dice and therefore the required expectation will be given by

$$E(x)=\sum xp(x)=(1+2+3+4+5+6)\frac{1}{6}=\frac{7}{2}.$$

3.2.3 Moment generating function

The M.g.f. of a probability function with parameter t is defined and denoted as

$$M_x(t) = E(e^{tx}) = \sum e^{x_i t} p_i(x) \quad \text{or} \quad \int_{-\infty}^{\infty} e^{tx} f(x)dx \tag{23}$$

for discrete and continuous case respectively.

If there are two independent random variables X and Y then

$$M_{X+Y}(t) = E(e^{tx+ty}) = E(e^{tx}).E(e^{ty}) = M_X(t).M_Y(t)$$

Also,
$$E(e^{tx_i}) = \sum f_i(1 + tx_i + \frac{t^2 x_i^{\,2}}{2!} + \frac{t^3 x_i^{\,3}}{3!} + ...)$$

$$= 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \frac{t^3}{3!}\mu_3' + ... + \frac{t^n}{n!}\mu_n' + ... \tag{24}$$

This implies in the expansion of moment generating function coefficient of t is μ_1',

coefficient of $\frac{t^2}{2!}$ is μ_2' and coefficient of $\frac{t^3}{3!}$ is μ_3'

Moreover, $\nu_1 = \left(\frac{d}{dt} M_x(t)\right)_{t=0}$ $\nu_2 = \left(\frac{d^2}{dt^2} M_x(t)\right)_{t=0}$ (25)

Example 11: Find the moment generating function of the following Uniform distribution

$f(x) = \frac{1}{b-a}$, $a < x < b$. Also find mean and variance.

$$M_x(t) = \int_a^b e^{tx} \frac{1}{b-a} dx = \frac{e^{bt} - e^{at}}{t(b-a)}$$

Now,
$$E(x) = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2} = mean$$

$$E(x^2) = \int_a^b x^2 \frac{1}{b-a} dx = \frac{a^2 + b^2 + ab}{3} \text{ and}$$

$$\text{Variance} = \frac{a^2 + b^2 + ab}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

Example 12: Find the moment generating function of $f(x) = pq^x, x = 0,1,2...$

Here $M_x(t) = E(e^{tx}) = \sum_{x=0}^{\infty} p(qe^t)^x = p(1+qe^t+(qe^t)^2+\ldots) = \frac{p}{1-qe^t}$

Example 13: Obtain the moment generating function of the random variable x having the following density function 2

$$f(x) = \begin{cases} x, 0 < x < 1 \\ 2-x, 1 < x < 2 \\ 0, x > 2 \end{cases}$$ Also determine ν_1, ν_2

$$M_x(t) = E(e^{tx}) = \int_0^1 xe^{tx}dx + \int_1^2 (2-x)e^{tx}dx$$

$$= \frac{e^{2t}-2e^t+1}{t^2} = \left(\frac{e^t-1}{t}\right)^2 = 1+t+t^2+\ldots$$

$$\nu_1 = \left(\frac{d}{dt}M_x(t)\right)_{t=0} = 1 \qquad \nu_2 = \left(\frac{d^2}{dt^2}M_x(t)\right)_{t=0} = 2$$

Problems for practice

1. Find the expectation from the following table

x	0	1	2	3	4
P(x)	1/4	1/4	0	1/4	1/4

   Ans: 2
2. Find the moment generating function for the continuous normal distribution given by $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; -\infty < x < \infty$ Ans $e^{\mu t + \frac{1}{2}t^2\sigma^2}$
3. Show that following function represents a provability density function $f(x) = pq^{x-1}, x = 1,2,3\ldots$
4. Find the moment generating function of the following Binomial distribution function $f(x) = C_x^n p^x q^{n-x}; x = 0,1,2\ldots$ also obtain mean and variance.

   Ans: $(p+pe^t)^n$, mean = np and variance = npq.

3.3 Method of least square

Let we want to fit a polynomial with n set of observations $(x_1, y_1),\ (x_2, y_2), \ldots (x_n, y_n)$. If $Y = f(x)$ be the exact curve then $Y - y$ represents the deviation or error say ε. Then the procedure to minimize the sum of squares of errors is known as method of least square.

We get normal equation, when differentiate the function with respect to unknowns. After solving these normal equations for required parameters we get the curve of the best fit of the given data.

Curve fitting of straight line

We have the error $\varepsilon_i = y_i - (a + bx_i) \Rightarrow \varepsilon_i^{\,2} = \sum [y_i - (a + bx_i)]^2$ (1)

Now we minimize $\varepsilon_i^{\,2}$. So differentiating w.r.t. a and b to get following normal equations

$$2\sum [y_i - (a + x_i)](-1) = 0 \Rightarrow \sum y_i = na + b\sum x_i \quad \text{and} \tag{2}$$

$$2\sum [y_i - (a + x_i)](-x_i) = 0 \Rightarrow \sum x_i y_i = a\sum x_i + b\sum x_i^{\,2}. \tag{3}$$

Example 14: Fit a straight line for the following data

y	12	15	21	25
x	50	70	100	120

The normal equations for the straight line are

$$\sum y_i = na + b\sum x_i \quad \text{and} \quad \sum x_i y_i = a\sum x_i + b\sum x_i^{\,2}.$$

Here, $n = 4,\ \sum y = 73,\ \sum x = 340,\ \sum xy = 6750$ and $\sum x^2 = 31800.$

Equations to be solved for a, b are

$$73 = 4a + 340b \quad \text{and} \quad 6750 = 340a + 31800b$$

which gives $a = 2.278,\ b = 0.187$.

Hence, required straight line is $y = 2.278 + 0.187x.$

Fitting of a second degree parabola $y = a + bx + cx^2$

Define the error term $\varepsilon_i = y_i - (a + bx_i + cx_i^{\,2})$ which implies

$$\varepsilon_i^{\,2} = \sum [y_i - (a + bx_i + cx_i^{\,2})]^2$$

Differentiating w.r.t. a, b, c respectively we get

$$2\sum[y_i-(a+bx_i+cx_i^{\,2})](-1)=0\Rightarrow\sum y_i=na+b\sum x_i+c\sum x_i^{\,2} \quad (4)$$

$$2\sum[y_i-(a+bx_i+cx_i^{\,2})](-x_i)=0\Rightarrow\sum x_iy_i=a\sum x_i+b\sum x_i^{\,2}+c\sum x_i^{\,3} \quad (5)$$

and

$$2\sum[y_i-(a+bx_i+cx_i^{\,2})](-x_i^{\,2})=0\Rightarrow\sum x_i^{\,2}y_i=a\sum x_i^{\,2}+b\sum x_i^{\,3}+c\sum x_i^{\,4} \quad (6)$$

Example 15: Fit a second degree parabola to the following data.

x	1	2	3	4
y	6	11	18	27

Since normal equations for for parabola are

$$\sum y_i=na+b\sum x_i+c\sum x_i^{\,2} \quad \sum x_iy_i=a\sum x_i+b\sum x_i^{\,2}+c\sum x_i^{\,3} \quad \text{and}$$
$$\sum x_i^{\,2}y_i=a\sum x_i^{\,2}+b\sum x_i^{\,3}+c\sum x_i^{\,4}$$

x	y	x^2	x^3	x^4	xy	x^2y
1	6	1	1	1	6	6
2	11	4	8	16	22	44
3	18	9	27	81	54	162
4	27	16	64	256	108	432
$\sum x=10$	$\sum y=62$	$\sum x^2=30$	$\sum x^3=100$	$\sum x^4=354$	$\sum xy=190$	$\sum x^2y=644$

Now, solving

$$62=4a+10b+30c,\ 190=10a+30b+100c \quad \text{and}$$

$$644=30a+100b+354c$$

We get $a=3,\ b=2$ and $c=1.$

And therefore $y=3+2x+x^2$ is the required parabola.

Fitting of an exponential curve $y=ae^{bx}$

Taking log both sides to get

$\log_{10} y=\log_{10} a+bx$ which is a straight line $Y=A+Bx$

With the similar process we find the normal equations for this straight line and determined a, b as follows

$$\sum Y = nA + B\sum x \quad \text{and} \quad \sum xY = A\sum x + B\sum x^2 \quad \text{where}$$

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

Example 16: For the following data fit the curve $y = ae^{bx}$

x	1	5	7	9	12
y	10	15	12	15	21

x	y	$Y = \log_{10} y$	x^2	xY
1	10	1.0	1	1
5	15	1.1761	25	5.8805
7	12	1.0792	49	7.5544
9	15	1.1761	81	10.5849
12	21	1.3222	144	15.8664
$\sum x = 34$		$\sum Y = 5.7536$	$\sum x^2 = 300$	$\sum xY = 40.8862$

The normal equations of exponential the curve are

$$\sum Y = nA + B\sum x \quad \text{and} \quad \sum xY = A\sum x + B\sum x^2$$

Using tabulated data, these equations become

$$5.7536 = 5A + 34B \quad \text{and} \quad 40.8862 = 34A + 300B$$

Solving which yields

$$A = 0.9766 \Rightarrow a = e^A = 9.475 \quad \text{and} \quad b = \frac{B}{\log_{10} e} = 0.059$$

Hence, $y = 9.475e^{0.059x}$ is the desired curve.

Example 17: Use the method of least square to fit the curve $y = \frac{c_0}{x} + c_1\sqrt{x}$ to the following

x	0.1	0.2	0.4	0.5	1	2
y	21	11	7	6	5	6

First we drive the normal equations with the usual process

$$\sum \frac{y}{x} = c_0\sum \frac{1}{x^2} + c_1\sum \frac{1}{\sqrt{x}} \quad \text{and} \quad \sum y\sqrt{x} = c_0\sum \frac{1}{\sqrt{x}} + c_1\sum x$$

Now we compute the following

$$\sum x = 4.2\,,\quad \sum \frac{y}{x} = 302.5,\quad \sum y\sqrt{x} = 33.715,\quad \sum \frac{1}{\sqrt{x}} = 10.1$$

and $\sum \frac{1}{x^2} = 136.5$

Solving the normal equations with the computed data we get,

$$c_0 = 1.977 \text{ and } c_1 = 3.281$$

Therefore, $y = \frac{1.977}{x} + 3.281\sqrt{x}$ is the desired curve.

Problems for practice

1. The pressure and volume of gas are related by the equation $pV^a = b$ where a, b are constants. Fit this equation to the following set of data:

p	0.5	1	1.5	2	2.5	3
V	1.62	1	0.75	0.62	0.52	0.46

Ans: $pV^{1.42} = 0.99.$

2. For the data given below, find the equation to the best fitting exponential curve of the form $y = ae^{bx}$

x	1	2	3	4	5
y	1.6	4.5	13.8	40.2	125

Ans: $y = 0.558e^{1.063x}$

3. Fit a curve of the type $xy = ax + b$ to the following data:

x	1	3	5	7	9	10
y	36	29	28	26	24	15

Ans: $xy = 16.18x + 40.78$

4. Fit a second degree parabola to the following data taking y as the dependent variable

x	1	2	3	4	5	6	7	8	9
Y	2	6	7	8	10	11	11	10	9

Ans: $y = -1 + 3.55x - 0.27x^2$

3.4 Correlation

In a bivariate distribution if the change in one variable affects the corresponding change in other variable, then both variables are called correlated. And the strength of their correlation is determined by correlation coefficient.

Mathematically the coefficient of correlation between two variables x, y is defined and

denoted by

$$r_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \quad (1)$$

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

where n is the number of observations, $\bar{x}$, $\bar{y}$ are the means of x and y respectively.

σ_x, σ_y are the standard deviations of x and y respectively and $Cov(x,y)$ represents the covariance of x and y.

Note: 1. $-1 \le r \le 1$.

2. Correlation coefficient r is independent of shifting and scaling both.

Example 18: Find the coefficient of correlation between the following values of x and y

x	y	x^2	y^2	xy
1	8	1	64	8
3	12	9	144	36
5	15	25	225	75
7	17	49	289	119
8	18	64	324	144
10	20	100	400	200
Total = 34	Total = 90	Total = 248	Total = 1446	Total = 582

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

$$= \frac{6(582) - 34(90)}{\sqrt{6(248) - (34)^2}\sqrt{6(1446) - (90)^2}}$$

On solving we get, correlation coefficient r = 0.9879.

Example 19: The variates x and y having mean zero are correlated with coefficient r. Two other variables u, v are defined by the relations $u = x\cos\alpha + y\sin\alpha$, $v = y\cos\alpha - x\sin\alpha$. Show that u, v will be uncorrelated if $\tan 2\alpha = \frac{2r\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}$, where σ_x, σ_y are the standard deviation of x and y respectively.

Given that $\bar{x} = \bar{y} = 0$, also

$u = x\cos\alpha + y\sin\alpha \Rightarrow \bar{u} = \bar{x}\cos\alpha + \bar{y}\sin\alpha$ and

$$v = y\cos\alpha - x\sin\alpha. \Rightarrow \bar{v} = \bar{y}\cos\alpha - \bar{x}\sin\alpha$$

Now, u, v will be uncorrelated implies $E(u-\bar{u})(v-\bar{v}) = 0$ (2)

$$\Rightarrow E[(x-\bar{x})\cos\alpha + (y-\bar{y})\sin\alpha][(y-\bar{y})\cos\alpha - (x-\bar{x})\sin\alpha] = 0$$

Or $\frac{1}{2}\sin 2\alpha[E(y^2) - E(x^2)] + \cos 2\alpha E(xy) = 0$

Or $\frac{1}{2}\sin 2\alpha[\sigma_x{}^2 - \sigma_y{}^2] + r.\cos 2\alpha.\sigma_x\sigma_y = 0$

Which implies, $\tan 2\alpha = \frac{2r\sigma_x\sigma_y}{\sigma_x{}^2 - \sigma_y{}^2}$.

Example 20: Find the coefficient of correlation of the following data

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

Here we will find the correlation coefficient with the shifting and scaling property.

x	y	$u = \frac{x-32}{4}$	$u = \frac{y-24}{6}$	u^2	v^2	uv
10	18	-3	-1	9	1	3
14	12	-2	-2	4	4	4
18	24	-1	0	1	0	0
22	6	0	-3	0	9	0
26	30	1	1	1	1	1
30	36	2	2	4	4	4
Total		-3	-3	19	19	12

$$r = \frac{n\sum uv - \sum u.\sum v}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}} = \frac{63}{105} = 0.6$$

Rank correlation for the given data is determined by the following formula

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (3)$$

where d_i is the difference of corresponding ranks and n is the number of observations.

If rank ties then the factor $\frac{m(m^2-1)}{12}$ is added in the numerator for each tied observation

having multiplicity m.

Example 21: Compute the rank correlation for the following data:

X	10	15	12	17	13	16	24	14	22
Y	30	42	45	46	33	34	40	35	39

Here ,

Rank (X)	9	5	8	3	7	4	1	6	2
Rank (Y)	9	3	2	1	8	7	4	6	5
d_i	0	2	6	2	-1	-3	-3	0	-3
d_i^2	0	4	36	4	1	9	9	0	9

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(72)}{9(81-1)} = 0.4$$

Example 22: Compute the rank correlation for the following data:

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

Clearly,

Rank (X)	4	6	2.5	9	6	1	2.5	10	8	6
Rank (Y)	5	7	3.5	10	1	6	3.5	9	8	2
d_i	-1	-1	-1	-1	5	-5	-1	1	0	4
d_i^2	1	1	1	1	25	25	1	1	0	16

Here 75, 64 and 68 occur two, three and two respectively.

Therefore rank correlation

$$\rho = 1 - \frac{6\left(\sum d_i^2 + \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} + \frac{m_3(m_3^2-1)}{12}\right)}{n(n^2-1)} \quad (4)$$

$$= 1 - \frac{6[72(12) + 2(3) + 3(8) + 2(3)}{10(99)(12)} = 0.54$$

Problems for practice

1. Calculate the coefficient of correlation for the following data:

x	65	66	67	67	68	69	70
y	67	68	65	68	72	72	69

Ans: 0.603

2. Find the correlation coefficient between x and y from the following data

x	60	34	40	50	45	41	22	43
y	75	32	34	40	45	33	12	30

Ans: 0.915.

3. If z = ax+by and r is the correlation coefficient between x and y, show that $\sigma_z^{\,2} = a^2\sigma_x^{\,2} + b^2\sigma_y^{\,2} + 2abr\sigma_x\sigma_y$.

4. The marks secured by recruits in the selection test (X) and in the proficiency text (Y) are given below:

X	10	15	12	17	13	16	24	14	22
Y	30	42	45	46	33	34	40	35	39

Find the rank correlation coefficient. Ans: 0.4

3.5 Regression

A regression line is a line having smallest deviation points to the data for best fit. We estimate y for given x with minimum possible error through regression line y on x and similarly we estimate x by y through regression line x on y.

For linear regression line we have two lines $y = a + bx$ or $x = a + by$

We know that for $y = a + bx$ the normal equations are

$$\sum y = na + b\sum x \quad \text{and} \quad \sum xy = a\sum x + b\sum x^2$$

Solving both equations for b we get regression coefficient

$$b_{yx} = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - \left(\sum x\right)^2} \qquad (1)$$

Similarly if we work with the line $x = a + by$ the other regression coefficient will be

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

So that the following results are obtained

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}, \quad b_{yx} = r\frac{\sigma_y}{\sigma_x} \tag{3}$$

and $$b_{xy}.b_{yx} = r^2 . \tag{4}$$

Regression coefficients b_{xy}, b_{yx} represents the slope of the line x on y and y on x respectively. Equation of regression line passes through mean point $(\bar{x}, \bar{y})$ having slope b_{yx} is

$$y - \bar{y} = b_{yx}(x - \bar{x}) . \tag{5}$$

Similarly equation of regression line passes through mean point $(\bar{x}, \bar{y})$ having slope b_{xy} is

$$x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow (y - \bar{y}) = \frac{1}{b_{xy}}(x - \bar{x}). \tag{6}$$

If θ be the angle between the regression lines having both slopes y on x as

$$\frac{\sigma_y}{r\sigma_x} \text{ and } r\frac{\sigma_y}{\sigma_x} \tag{7}$$

then $$\tan\theta = \left|\frac{b_{xy} - b_{yx}}{1 + b_{xy}b_{yx}}\right| = \left|\frac{\frac{r\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}}{1 + \frac{r\sigma_y}{\sigma_x}.\frac{\sigma_y}{r\sigma_x}}\right| = \frac{(1-r^2)}{r}\frac{\sigma_x\sigma_y}{\sigma_x{}^2 + \sigma_y{}^2}. \tag{8}$$

Now consider the following cases:

Case I- when $r = 0 \Rightarrow \tan\theta = \infty \Rightarrow \theta = \pi/2$ that is both the regression line are perpendicular.

Case II- when $r = \pm 1 \Rightarrow \tan\theta = 0 \Rightarrow \theta = 0$ that is both regression lines coincide.

We have also following points for the consideration:

1. At a time both the regression coefficients have same sign.
2. Since $b_{xy}.b_{yx} = r^2 \le 1$ this implies both regression coefficients cannot be greater than 1
   simultaneously. However, if one of them is greater than 1, other must be less than 1.

3. Both lines of regression always passes through mean point $(\bar{x}, \bar{y})$.

Example 23: From the regression lines $3x = 1 + 4y$ and $2x - y = 2$ find mean of x and y. Hence determine correlation coefficient r.

Solving both equations for x and y we get the mean point $(\bar{x}, \bar{y}) = \left(\frac{7}{5}, \frac{4}{5}\right)$.

Also we may write the equations as $x = \frac{1}{2}y + 1$ and $y = \frac{3}{4}x - \frac{1}{4}$

Therefore, $b_{xy} = \frac{1}{2}, b_{yx} = \frac{3}{4}$.

Again $b_{xy}.b_{yx} = r^2 \Rightarrow r = \sqrt{b_{xy}b_{yx}} = \sqrt{\frac{3}{8}}$

Example 24: From the following data of 10 observations find the lines of regression.

$$\sum xy = 80, \ \sum x = 24, \ \sum y = 30, \quad \sum x^2 = 66, \quad \sum y^2 = 110,$$

Since $b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} = 0.952$ and $b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2} = 0.4$

and $r = \sqrt{b_{xy}b_{yx}} = \sqrt{0.952(4)} = 0.617$

$$\bar{x} = \frac{\sum x}{n} = \frac{24}{10} = 2.4 \text{ and } \bar{y} = \frac{\sum y}{n} = \frac{30}{10} = 3$$

Now regression lines are determined by the following

$$y - \bar{y} = b_{yx}(x - \bar{x}) \Rightarrow y - 3 = 0.952(x - 2.4) \text{ and}$$

$$x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow x - 2.4 = 0.4(y - 3).$$

Example 25: The following results were obtained in the analysis of data on yield of drug bank in ounce (y) and age in years (x) of 200 plants:

	x	y
Average	9.2	16.5
Standard deviation	2.1	4.2

Construct the two lines of regression and estimate the yield of a plant of age 8 years.

Correlation coefficient is given 0.84.

We have $\bar{x} = 9.2, \ \bar{y} = 16.5, \quad r = 0.84, \quad \sigma_x = 2.1, \quad \text{and} \quad \sigma_y = 4.2.$

First we find regression coefficients

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.42, \quad b_{yx} = r\frac{\sigma_y}{\sigma_x} = 1.68.$$

Now, regression lines are

$$y - \bar{y} = b_{yx}(x - \bar{x}) \Rightarrow y = 1.68x + 1.044$$

and

$$x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow x = 0.42y + 2.27 .$$

Also, $y(8) = 14.484.$

3.6 Multiple linear regression

If the data has more than two variables then the regression is called multi linear regression.

Suppose we have to find a regression plane $y = a + bx + cz$ through the data points of x, y, z .

Now the error term is $\varepsilon = [y - (a + bx + cz)].$

With method of least square, we minimize the sum of square of errors

$$\varepsilon_i^2 = \sum [y - (a + bx + cz)]^2 .$$

Differentiating w.r.t. a, b and c we get

$$2\sum [y - (a + bx + cz)](-1) = 0 \Rightarrow \sum y = na + b\sum x + c\sum z \qquad (1)$$

$$2\sum [y - (a + bx + cz)](-x) = 0 \Rightarrow \sum xy = a\sum x + b\sum x^2 + c\sum zx, \qquad (2)$$

and

$$2\sum [y - (a + bx + cz)](-1) = 0 \Rightarrow \sum zy = a\sum z + b\sum zx + c\sum z^2 \qquad (3)$$

respectively.

These three equations are known as normal equation to fit the desired plane.

Example 26: From the following data, find the multiple linear regression x on y and z

X	1	3	0	2
Y	2	1	1	3

Z	2	2	0	1

We compute the observations as

$$\sum x = 5,\ \sum y = 7,\ \sum z = 5,\ \sum y^2 = 15,\ \sum xy = 11,\ \sum yz = 9,$$

$$\sum xz = 10,\ \text{and}\ \sum z^2 = 9.$$

Now, the normal equations for the plane $x = a + by + cz$ are

$$\sum x = 4a + b\sum y + c\sum z$$

$$\sum xy = a\sum y + b\sum y^2 + c\sum yz,\ \text{and}$$

$$\sum xz = a\sum z + b\sum zy + c\sum z^2.$$

These equations are

$5 = 4a + 7b + 5c,\ 11 = 7a + 15b + 9c,$ and $10 = 5a + 9b + 9c$ respectively

Solving for a, b, c we get $a = -\frac{8}{5},\ b = \frac{7}{10}$ and $c = \frac{13}{10}.$

Hence, $10x = 7y + 13z - 16$ is the required regression line.

Problems for practice

1. Find both the lines of regression of the following data:

x	5.6	5.65	5.7	5.81	5.85
y	5.8	5.7	5.8	5.79	6.01

   Ans: y =0.743x+1.568 and x = 0.636y+2.02

2. Two lines of regression are given by x+2y-5 =0 and 2x+3y-8 =0 and $\sigma_x^2 = 12$. Find (i) the mean values of x and y (ii) variance of y (iii) The coefficient of correlation

   Ans: $\bar{x} = 1, \bar{y} = 2,$ variance (y) = 4 and $r = \frac{-\sqrt{3}}{2}$

3. The equations of two regression line obtained in a correlation analysis of 60 observations are: 5x = 6y+24 and 1000y=768x- 3608.
   What is the correlation coefficient? Show that the coefficient of variation of x and y is 5/24. Ans: 0.96

4. Using method of least square fit the plane Z = a+bX+cY for the following data

Y	1	2	3	4
X	0	1	2	3
Z	12	8	24	30

   Ans: Z = 10+4x+2Y

Links for e-content

(Central tendency)

https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/

http://www.learnalberta.ca/content/memg/Division03/Measures%20of%20Central%20Tendency/index.html

https://www.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap/measuring-center-quantitative/v/statistics-intro-mean-median-and-mode

https://www.youtube.com/watch?v=AdH5vfobH5E

(Moments, Skewness and Kurtosis)

https://spocathon.page/video/lecture-24-skewness-and-kurtosis

https://www.youtube.com/watch?v=m9a6rg0tNSM

https://www.youtube.com/watch?v=BsVtMnp3vks

https://www.youtube.com/watch?v=OfANWrzQE9Q

https://www.youtube.com/watch?v=1da4auXziT8

https://ocw.mit.edu/courses/brain-and-cognitive-sciences/9-07-statistics-for-brain-and-cognitive-science-fall-2016/lecture-notes/MIT9_07F16_lec6.pdf

(Curve fitting)

https://nptel.ac.in/courses/111/104/111104120/

https://www.khanacademy.org/math/linear-algebra/alternate-bases/orthogonal-projections/v/linear-algebra-least-squares-approximation

https://theengineeringmaths.com/wp-content/uploads/2018/01/curve-fitting-and-correlation.pdf

https://www.accessengineeringlibrary.com/content/book/9780071795579/chapter/chapter8

https://www.youtube.com/watch?v=FmGorE6f1ik

https://www.youtube.com/watch?v=vKRwnc4SzbU

https://www.youtube.com/watch?v=AzroLr1XS5E

(Correlation)

https://www.accessengineeringlibrary.com/content/book/9780071795579/chapter/chapter8

https://theengineeringmaths.com/wp-content/uploads/2018/01/curve-fitting-and-correlation.pdf

https://www.youtube.com/watch?v=xUwnB3RrDAw

https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data#scatterplots-and-correlation

https://www.youtube.com/watch?v=vv-l0vOayKM

(Regression analysis)

https://www.youtube.com/watch?v=uuxSKqU-qBY

http://www.cimt.org.uk/projects/mepres/alevel/stats_ch12.pdf

https://www.coconino.edu/resources/files/pdfs/academics/sabbatical-reports/kate-kozak/chapter_10.pdf

https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data#regression-library

https://www.accessengineeringlibrary.com/content/book/9780071795579/chapter/chapter8

https://www.youtube.com/watch?v=QAEZOhE13Wg