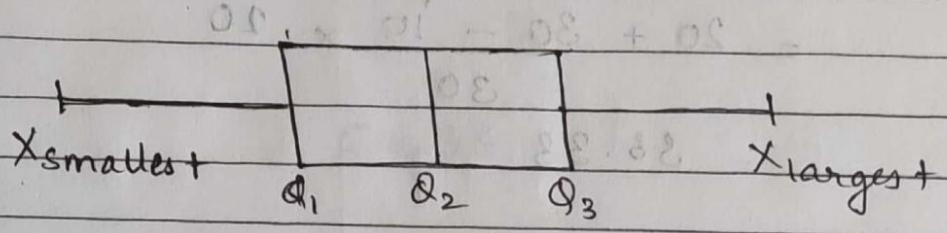
3) Box Plot (five number summary)

The number that represents smallest value ($X_{smallest}$), first quartile (q_1), second quartile (q_2), third quartile (q_3) and the largest value ($X_{largest}$) is called five number summary.

The graphical representation of five number summary in a box connected with straight line is called box plot.



From the following data, construct a box plot:

Marks	0-20	20-40	40-60	60-80	80-100
No. of students	10	30	36	30	14

Sol^n: Here,

Marks	No. of students (f)	c.f
0-20	10	10
20-40	30	40
40-60	36	76
60-80	30	106
80-100	14	120

for Q_1:

$$\frac{N}{4} = \frac{120}{4} = 30$$

So, Q_1 lies in class 20-40

$\therefore L = 20, f = 30, c.f = 10$

Then,

$$Q_1 = L + \frac{\frac{N}{4} \times 1 - c.f}{f} \times h$$

$$= 20 + \frac{30 - 10}{30} \times 20$$

$$= 33.33$$

for Q_2:

$$\frac{2N}{4} = \frac{N}{2} = 60$$

So, Q_2 lies in class 40-60.

$\therefore L = 40, c.f = 40, f = 36$

Then,

$$Q_2 = L + \frac{N/4 \times 2 - c.f}{f} \times h$$

$$= 40 + \frac{60 - 40}{36} \times 20$$

$$= 51.11$$

for Q_3 :

$$Q_3 = \frac{3N}{4} = \frac{3 \times 120}{4} = 90$$

So Q_3 lies in class 60-80.

$\therefore L = 60,\ f = 30,\ cf = 76.$

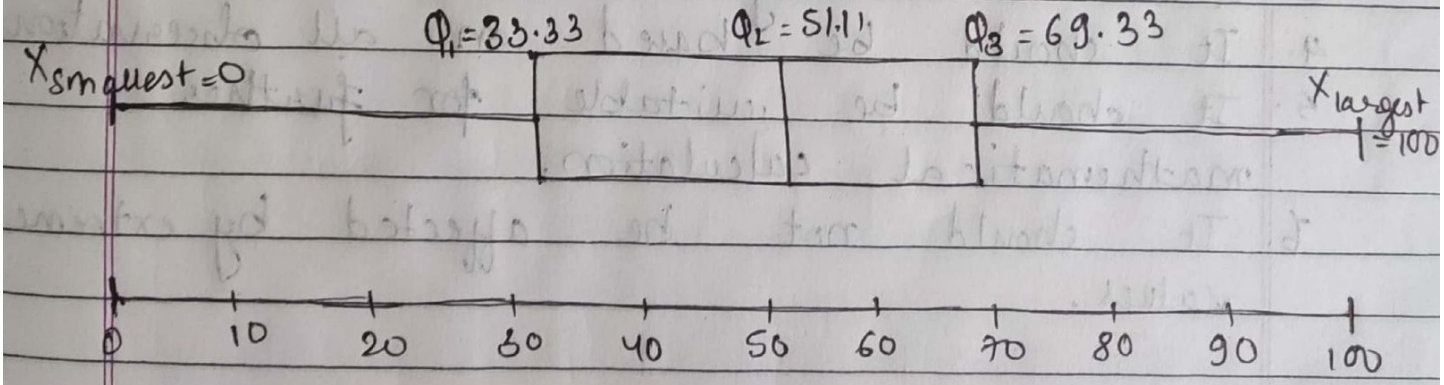
Then,

$$Q_3 = L + \frac{\frac{3N}{4} - c.f}{f} \times h$$

$$= 60 + \frac{90 - 76}{30} \times 20$$

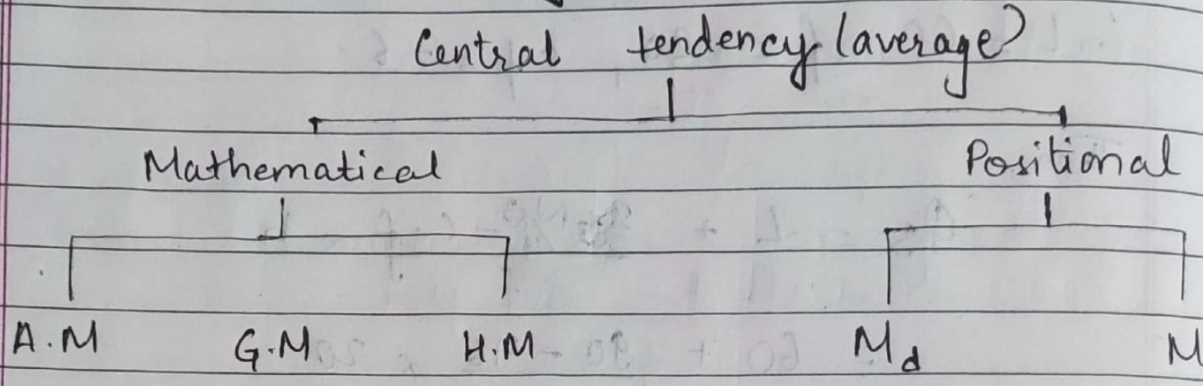
$$= 69.33$$

Now, the box plot is :



- Measure of central tendency.

The measure of central value or the average of the given data is called measure of central tendency.



- Criteria for good average:

1. It should be rigidly defined.
2. It should be easy to calculate.
3. It should be easy to understand.
4. It should be based on all observations.
5. It should be suitable for further mathematical calculation.
6. It should not be affected by extreme values.

- Dispersion:

The measure of scattering of data is called dispersion. It is also called variation. Dispersion can be measured in two ways:

1. Absolute measure of dispersion:

If the dispersion value is expressed in the original unit of given data, then it is called absolute measure of dispersion. Eg: ~~Range~~ Range, standard deviation, mean deviation, quartile deviation, etc.

2. Relative measure of dispersion:

If the dispersion value is independent of original unit of given data, then it is called relative measure of dispersion. Eg: Coefficient of range, coefficient of SD, coeff. of M.D, coefficient of Q.D, etc.

Some formula:

1. Range = Largest − smallest value = L − S
2. Interquartile range = $Q_3 - Q_1$
3. Quartile deviation = $\frac{Q_3 - Q_1}{2}$
4. Standard deviation (σ) = $\sqrt{\frac{\Sigma f x^2}{N} - \left(\frac{\Sigma f x}{N}\right)^2}$

   or,

   $\sigma = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$
5. Variance = σ^2
6. Mean deviation from $\bar{x}$ = $\frac{1}{N} \Sigma f |x - \bar{x}|$
7. Coeff. of range = $\frac{L - S}{L + S}$
8. Coeff. of Q.D = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$
9. Coeff. of S.D = $\frac{\sigma}{\mu}$
10. Coeff. of variation = $\frac{\sigma}{\mu} \times 100\%$.

11. Coeff. of M.D from $\bar{x}$ = $\frac{\text{M.D from } \bar{x}}{\bar{x}}$

12. $Q_i = L + \frac{iN/4 - c.f}{f} \times h$

or,

value of $\left(\frac{n+1}{4}\right)^{i}$ th value

13. $D_i = L + \frac{iN/10 - c.f}{f} \times h$

or

value of $\frac{(n+1)}{10} \times i^{th}$ value

14 $P_i = L + \frac{iN/100 - c.f}{f} \times 100$

or,

$\left(\frac{n+1}{100}\right) i^{th}$ value.

15. Combine mean $\bar{x}_{123} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3}$

16. Combine S.D $\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$

where,

$d_1 = \bar{x}_1 - \bar{x}_{12}$

$d_2 = \bar{x}_2 - \bar{x}_{12}$

Q. Find the missing frequency from the data:

Marks	0-20	20-40	40-60	60-80	80-100
No. of stu.	10	–	30	–	14

Given, $N = 94$ and mode $= 54$.

Sol: Here, let the missing frequency be x & y for the class interval 20-40 & 60-80 respectively.

Marks	No. of students (f)
0-20	10
20-40	x
40-60	30
60-80	y
80-100	14

Here,

$$f_0 = 30, \quad f_m = x, \quad f_2 = y$$

Then,

$$M_o = L + \frac{f_m - f_0}{2f_m - f_2 - f_0} \times h$$

$$\text{or, } 54 = 40 + \frac{30 - x}{30 \times 2 - y - x} \times 20$$

$$\text{or, } 14 = \frac{600 - 20x}{60 - y - x}$$

$$\text{or, } 840 - 14x - 14y = 600 - 20x$$

$$\Rightarrow -6x + 14y = 240 \quad \text{— (1)}$$

Also,

$10 + x + 30 + y + 14 = 94$

$\Rightarrow x + y = 40 \quad \text{---(ii)}$

$\therefore x = 16$

$y = 24$

Q. An investor buys Rs. 12000 worth of shares in a company each month. During the first five months, he bought the share at price of Rs. 100, 120, 150, 200, 240 per share. After five months, what is average price paid for share by him?

Sol^n: Here,

Price	Share per month ~~Average~~	$f(x)$
100	120	12,000
120	100	12,000
150	80	12,000
200	60	12,000
240	50	12,000
	$N = 410$	$\Sigma fx = 60,000$

$\therefore$ Average $= \dfrac{12000 \times 5}{410}$

$= 146.34$

Q An analysis of monthly wages paid to workers in two firms A & B

1) which firm A or B has large wage bill?
2) which firm has greater variability in wages
3) Calculate combined mean & variance.

	firm A	Firm B
No. of workers	500	600
Averages	186	175
Variance	81	100

Sol?: Here,

$n_1 = 500$; $\bar{x}_1 = 186$; $\sigma_1^2 = 81$

$n_2 = 600$; $\bar{x}_2 = 175$; $\sigma_2^2 = 100$

1) Since, $\bar{x} = \frac{\Sigma x}{n} \Rightarrow \Sigma x = n\bar{x}$

$\therefore \Sigma x_1 = n_1 \bar{x}_1 = 500 \times 186 = 93000$

$\Sigma x_2 = n_2 \bar{x}_2 = 600 \times 175 = 105000$

Here, $\Sigma x_2 > \Sigma x_1$, so firm B has large wage bill.

2) $CV_1 = \frac{\sigma_1}{\bar{x}_1} \times 100\% = \frac{9}{186} \times 100\% = 4.838\%$

$CV_2 = \frac{\sigma_2}{\bar{x}_2} \times 100\% = \frac{10}{175} \times 100\% = 5.714\%$

Here, $CV_2 > CV_1$, so B has greater variability.

3) $$\bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = 180$$

$$\sigma_{12}^2 = \frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}$$

where,

$$d_1 = \bar{x}_1 - \bar{x}_{12}$$

$$d_2 = \bar{x}_2 - \bar{x}_{12}$$

Correlation and Regression.

Correlation:

The statistical measure that is used to find the degree of relationship between two or more variable is called correlation and the value that gives the degree of relationship is called correlation coefficient (r).

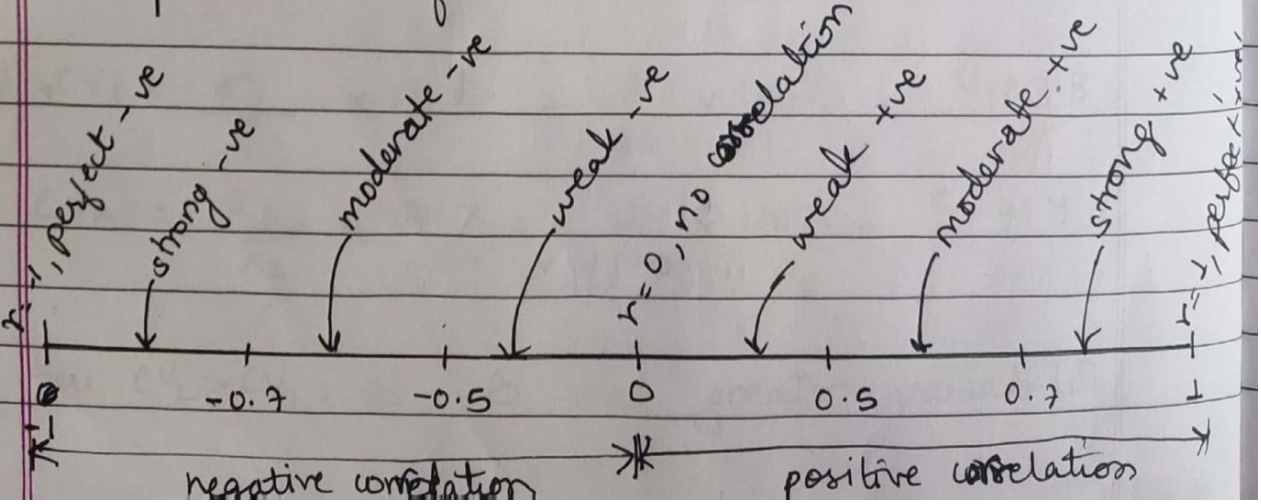
where,

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

- Properties of correlation coefficient:

1. It has no unit
2. It gives the degree of relationship between the variable.
3. The range of r is: $-1 \leq r \leq 1$
4. It is the geometric mean between the regression coefficient i.e: $r = \pm\sqrt{b_{xy} \cdot b_{yx}}$.
5. It is independent of change of origin and scale.
6. $r_{xy} = r_{yx}$

- Interpretation of 'r':



- Coefficient of determination:

Let r be the corelation coefficient, then r^2 is called coefficient of determination.

∴ Coefficient of determination $= r^2$
Coefficient of non-determination $= 1 - r^2$

- Interpretation of coefficient of determination

Let $r = 0.3$ then, $r^2 = 0.09$

∴ 9% of variation in independent variable y is due to x. Remaining 91% is due to other factor.

- Multiple correlation:

When there are more than two variable the degree of relationship between a dependent variable and the joint effect of independent variable is called multiple correlation.

Let x_1, x_2, x_3 be three variable, then the correlation coefficient between x_1 and joint effect of x_2 and x_3 is:

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

Similarly,

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

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

- Properties of multiple correlation coefficient:

1. The range of multiple correlation coefficient is $[0, 1]$
2. $R_{1.23} = R_{1.32}$
   $R_{2.13} = R_{2.31}$
   $R_{3.12} = R_{3.21}$

- Coefficient of multiple determination:

The square of coefficient of multiple correlation is called multiple determination.

- Interpretation:

Let $R_{1.23} = 0.5 \Rightarrow R_{1.23}^2 = 0.25$

This means 25% of variation in x is due to the joint effect of x_2 and x_3. Remaining 75% is due to other cause.

- Partial correlation coefficient:

The simple correlation between two variables holding other as constant is called partial correlation coefficient.

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

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

$$r_{23.1} = \frac{1 - r_{23}^2}{\sqrt{1 - r_{21}^2}\sqrt{1 - r_{13}^2}}$$

- Properties of partial correlation coefficient:

1. It lies between -1 to 1.
2. $r_{12.3} = r_{21.3}$
   $r_{13.2} = r_{31.2}$
   $r_{23.1} = r_{32.1}$
3.

- Coefficient of partial determination.

Square of partial correlation coefficient is called coefficient of partial determination.

- Interpretation:

Let $r_{12.3}^2 = 0.9$

90 %. of variation in x_1 is due to x_2 holding x_3 as constant. Remaining 10 %. is due to other cause.

Examples:

Q1. Calculate the correlation coefficient and interpret it.

X	39	65	62	90	86	75	25	98	36	78
Y	47	53	58	86	62	68	60	91	51	84

Also, calculate coefficient of determination & interpret it.

Sol^n: Here,

Correlation coefficient $(r) = \dfrac{n\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

From the given data, we get,

$n = 10$

$\Sigma xy = 45852$

$\Sigma x = 654$

$\Sigma y = 660$

$\Sigma x^2 = 48320$

$\Sigma y^2 = 45784$

Then,

$$r = \frac{10 \times 45852 - 654 \times 660}{\sqrt{10 \times 48320 - (654)^2}\ \sqrt{10 \times 45784 - (660)^2}}$$

$$= 0.765$$

∴ There is a strong positive correlation between x and y.

Coefficient of determination $(r^2) = 0.585$

∴ 58.5% variation in dependent variable is due to independent variable. Remaining 41.5% due to other causes.

Q2 A sample of 10 values of variable x_1, x_2, x_3 were obtained as:

$\Sigma x_1 = 10$, $\Sigma x_2 = 20$, $\Sigma x_3 = 30$

$\Sigma x_1^2 = 20$, $\Sigma x_2^2 = 68$, $\Sigma x_3^2 = 170$

$\Sigma x_1 \cdot x_2 = 10$, $\Sigma x_1 x_3 = 15$, $\Sigma x_2 x_3 = 64$

find:

a. Correlation coefficient between x_1 and x_2 holding x_3 as constant ($r_{12.3}$)

b. Correlation coefficient between x_1, x_2 and x_3 assuming x_1 as dependent variable. ($R_{1.23}$)

Sol^n: Here,

$$\therefore r_{12} = \frac{n\Sigma x_1 x_2 - \Sigma x_1 \Sigma x_2}{\sqrt{n\Sigma x_1^2 - (\Sigma x_1)^2}\sqrt{n\Sigma x_2^2 - (\Sigma x_2)^2}}$$

$$= \frac{10 \times 10 - 10 \times 20}{\sqrt{10 \times 20 - (10)^2}\sqrt{10 \times 68 - (20)^2}}$$

$$= -0.597$$

Similarly,

$$r_{23} = \frac{n\Sigma x_2 x_3 - \Sigma x_2 \Sigma x_3}{\sqrt{n\Sigma x_2^2 - (\Sigma x_2)^2}\sqrt{n\Sigma x_3^2 - (\Sigma x_3)^2}}$$

$$= \frac{10 \times 64 - 20 \times 30}{\sqrt{10 \times 68 - (20)^2}\sqrt{10 \times 170 - (30)^2}}$$

$$= 0.084$$

And,

$$r_{13} = \frac{n\Sigma x_1 x_3 - \Sigma x_1 \Sigma x_3}{\sqrt{n\Sigma x_1^2 - (\Sigma x_1)^2}\sqrt{n\Sigma x_3^2 - (\Sigma x_3)^2}}$$

$$= \frac{10 \times 15 - 10 \times 30}{\sqrt{10 \times 20 - (10)^2} \times \sqrt{10 \times 170 - (30)^2}}$$

$$= -0.53$$

Then,

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

$$= \frac{1 + -(-0.597)^2}{\sqrt{1-(0.53)^2} \times \sqrt{1-(0.084)^2}}$$

$$= 0.761$$

Again, And,

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

$$= \sqrt{\frac{(+0.597)^2 + (-0.53)^2 - 2\times(-0.597)(0.53)(0.084)}{1-(0.084)^2}}$$

$$= 0.767$$

* Regression:

The technique that develops the mathe-matical equation that describes the relationship between two variable i.e dependent and independent is called regression.

* Regression coefficient:

The value that represents the rate of change in dependent variable due to unit change in independent variable x is called regression coefficient (b_{yx})

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

* Properties of regression coefficient:

i) It measures the rate of change in dependent variable for unit change in independent variable.

ii) Correlation coefficient is geometric mean of regression coefficient i.e $r = \pm\sqrt{b_{xy} \cdot b_{yx}}$

iii) Both regression coefficient have same sign.

iv) $b_{xy} \neq b_{yx}$

v) $\frac{b_{xy} + b_{yx}}{2} > r$

vi) Independent of change of origin but not of scale.

* Least Square method for fitting regression line.

The regression line of y on x i.e $y = a + bx$ ...(1) can be fitted by using normal eqns:

$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Here,

a = y-intercept (B_0)

b = regression coeff. of y on x = slope (B_1)

We know,

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

$$\Rightarrow b_{yx} = r\frac{\delta y}{\delta x}, \quad b_{xy} = r\frac{\delta x}{\delta y}$$

Multiple Regression:

A equation that express the relationship between a dependent variable Y and two or more independent variable $(x_i, i = 1, 2, \cdots, n)$ is called multiple regression.

If z is dependent & x & y are independent variable then the regression eqn is:

$$z = a + bx + cy$$

where,

a = intercept

b = regression coeff. of z on x keeping Y constant

c = regression coeff. of z on Y keeping x constant

The normal equation for regression plane $z = a + bx + cy$ is:

$$\Sigma z = na + b\Sigma x + c\Sigma y$$
$$\Sigma xz = a\Sigma x + b\Sigma x^2 + c\Sigma xy$$
$$\Sigma yz = a\Sigma y + b\Sigma xy + c\Sigma y^2.$$

Coefficient of multiple determination:

→ for line $x_1 = a + bx_2 + cx_3$

$$R_{1.23} = \sqrt{\frac{a\Sigma x_1 + b\Sigma x_1 x_2 + c\Sigma x_1 x_3 - n\bar{x}_1^2}{\Sigma x_1^2 - n\bar{x}_1^2}}$$

→ for $x_2 = a + bx_1 + cx_3$

$$R_{2.13} = \sqrt{\frac{a\Sigma x_2 + b\Sigma x_1 x_2 + c\Sigma x_2 x_3 - n\bar{x}_2^2}{\Sigma x_2^2 - n\bar{x}_2^2}}$$

→ for $x_3 = a + bx_1 + cx_2$

$$R_{3.12} = \sqrt{\frac{a\Sigma x_3 + b\Sigma x_1 x_3 + c\Sigma x_2 x_3 - n\bar{x}_3^2}{\Sigma x_3^2 - n\bar{x}_3^2}}$$

* Standard error of estimate for regression:

The standard error of estimate of Y on X is measure of dispersion of observed value around $y = a + bx$ is given by:

$$S_e = S_{yx} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}}$$

* Confidence interval:

Let $y = a + bx$ be regression line then $(1-\alpha)\,100\%$ C.I for:

(i) intercept (β_0) is

$(a - t_{\alpha/2, n-2} \cdot S_a ; a + t_{\alpha/2, n-2} S_a)$

where,

$$S_a = \frac{\sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}} \times \sqrt{\frac{\Sigma x^2}{n}}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}}$$

(ii) slope (β_1) is :

$(b - t_{\alpha/2, n-2} \cdot S_b , b + t_{\alpha/2, n-2} \cdot S_b)$,

where,

$$S_b = \frac{\sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}}$$

Q. The following are measurement of air velocity and evaporation coefficient of burning fuel droplet in impulse engine.

Air velocity (cm/sec)	20	60	100	140	180	220	260	300	340	380
evaporation coefficient (mm^2/sec)	0.18	0.37	0.35	0.78	0.56	0.75	1.18	1.36	1.17	1.65

a) fit the regression line to these data & use it to estimate the evaporation coefficient of a droplet when air velocity is 190 cm/sec

(Air velocity given, so assume it to be x)

b) Construct 95% C.I. for intercept & slope.

x - independent variable

Soln: Here,

let x represent air velocity and y represent the evaporation coefficient then from given data:

x	y
20	0.18
60	0.37
100	0.35
140	0.78
180	0.56
220	0.75
260	1.18
300	1.36
340	1.17
380	1.65

Then, from table:

$n = 10$, $\Sigma x = 2000$, $\Sigma y = 8.35$

$\Sigma xy = 2175.40$, $\Sigma x^2 = 532000$

$\Sigma y^2 = 9.1079$

Then,

the required regression eqn is $y = a + bx$ —(i)
with normal equation:

$$\Sigma y = na + b\Sigma x$$
$$\Rightarrow 8.35 = 10a + 2900b \quad \text{—(ii)}$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$
$$\Rightarrow 2175.40 = 2000a + 532000b \quad \text{—(iii)}$$

Solving (ii) & (iii);

$$a = 0.069$$
$$b = 0.00383$$

So, the required regression line is;

$$y = 0.069 + 0.00383x$$

At 190 cm/sec air velocity; evaporation coeff.

$$\Rightarrow y = 0.069 + 0.00383 \times 190$$
$$= 0.7967$$
$$\approx 0.80 \text{ mm}^2/\text{sec}$$

(190 ↑ in calculator: regression mode)

b) CI for intercept.

given $(1-\alpha)\,100\% = 95\%$.

$$\Rightarrow \alpha = 0.05$$

The 95% CI is :

$(a - t_{\alpha/2, n-2} \cdot S_a,\ a + t_{\alpha/2, n-2} \cdot S_a)$

$\Rightarrow (0.069 - t_{\frac{0.05}{2}, 10-2} \cdot S_a,\ 0.069 + t_{0.05/2, 10-2} \cdot S_a)$

We know:

$$S_a = \frac{\sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}} \cdot \sqrt{\Sigma x^2}}{\sqrt{\Sigma x^2 - (\Sigma x)^2/n}}$$

$$= \frac{\sqrt{\frac{9.1079 - 0.069 \times 8.35 - 0.00383 \times 2175.4}{10-2}} \times \sqrt{532000}}{\sqrt{532000 - \frac{(2000)^2}{10}}}$$

~~36.462~~

$$= 0.101$$

So C.I is:

$(0.069$

thorai incomplete

Q. The following table shows the weight z to the nearest pound, height x to the nearest inch, and age y to the nearest years, of 12 boys.

Weight(z)	64	71	53	67	55	58	77	57	56	51	76	68
Height(x)	57	59	49	62	51	50	55	48	52	42	61	57
Age(y)	8	10	6	11	8	7	10	9	10	6	12	9

(i) Fit the regression plane & estimate the weight of boy is 9 yrs old & 54 inches tall.

(ii) Fit the regression line & estimate the weight of boy who is 15 yrs old.

Sol^n. Here,

From the data given;

$\Sigma x = 643$, $\Sigma y = 106$

$\Sigma xy = 5779$, $\Sigma x^2 = 34843$

$\Sigma y^2 = 976$, $n = 12$, $\bar{x} = 53.5$, $\bar{y} = 8.83$

From z & y;

$\Sigma z = 753$, $\Sigma z^2 = 48139$

$\bar{z} = 62.435$, $\Sigma yz = 6796$

$\Sigma xz = 40830$

Let z represent weight, x represent height, y represent age, then the regression plane to fit is:

$$z = a + bx + cy \quad -(1) \text{ with normal eqn:}$$

$$\Sigma z = na + b\Sigma x + c\Sigma y$$
$$\Rightarrow 753 = 12a + 643b + 106c$$

$$\Sigma xz = a\Sigma x + b\Sigma x^2 + c\Sigma xy$$
$$\Rightarrow 40830 = 643a + 34843b + 5779c$$

$$\Sigma yz = a\Sigma y + b\Sigma xy + c\Sigma y^2$$
$$\Rightarrow 6796 = 106a + 5779b + 976c$$

On solving:

$$a = 3.65$$
$$b = 0.855$$
$$c = 1.506$$

So, the required regression plane is:

$$z = 3.65 + 0.855x + 1.506y$$

(i) At 9 years old & 54 inches tall;

$$z = 3.65 + 0.855 \times 54 + 1.506 \times 9$$
$$\Rightarrow z = 63.356 \simeq 63 \text{ , proved.}$$

At 15 yrs old;

(ii) $z = a + by$

$z = 3.65 + 0.855 \times 15$

$= 16.475$

Again;

(i) $R_{z \cdot xy} = \sqrt{\dfrac{a\Sigma z + b\Sigma xz + c\Sigma yz - n\bar{z}^2}{\Sigma z^2 - n\bar{z}^2}}$

$= \sqrt{\dfrac{3.65 \times 753 + 0.855 \times \frac{643 \times 753}{40830} + 1.506 \times 6796 - 12 \times (62.435)^2}{48139 - 12 \times (62.435)^2}}$

$= 0.850$

Chapter 8:

Application of Computer in statistical computing

- Some useful formula:

Sample mean $(\bar{x}) = \frac{\Sigma x}{n}$

Sample S.D $(s) = \sqrt{\frac{\Sigma x^2}{n-1} - \frac{n\bar{x}^2}{n-1}}$

Standard error of sample mean $SE(\bar{x}) = \frac{s}{\sqrt{n}}$

C.V of sample $= \frac{s}{\bar{x}} \times 100\ \%$

Sample variance $= s^2$

Population mean $\mu = \frac{\Sigma x}{N}$

Population S.D $(\sigma) = \sqrt{\frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2}$

C.I for sample mean : $\left(\bar{X} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right)$

Combined sample standard deviation

$$s_{12} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2 - 2}}$$

where,

$$d_1 = \bar{X}_1 - \bar{X}_{12}, \quad d_2 = \bar{X}_2 - \bar{X}_{12}$$

$$\bar{X}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Standard error of difference of mean

$$SE(\bar{x} - \bar{y}) = \sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}$$

C.I for sample variance:

$$\left(\frac{(n-1)\, s^2}{\chi^2_{\alpha/2,\, n-1}} \; ; \; \frac{(n-1)\, s^2}{\chi^2_{1-\alpha/2,\, n-1}} \right)$$

Standard deviation of difference of average:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

Q. Following data reveal the sample of 22 pairs of (x, y) drawn from large pop^n.

X	46	61	56	68	58	45	50	59	45	66	57	59	66	62
Y	49	46	43	82	26	27	29	47	37	30	43	32	27	37

	57	57	45	50	61	55	47	51
.	24	43	49	48	29	37	32	26

(i) find sample mean for each ~~x & y~~ X and Y.
(ii) find sample S.D for each X and Y.
(iii) which sample is good?
(iv) which series is consistent and why?
(v) find Karl Pearson correlation coefficient.
(vi) find standard error of difference of mean.
(vii) construct box plot for X
(viii) find 95% C.I for sample mean of X

Soln: Here,

i) from calculator

$\Sigma x = 1221$

$\Sigma y = 793$

$n = 22$

$\bar{x} = \frac{\Sigma x}{n} = \frac{1221}{22} = 55.5$

$\bar{y} = \frac{\Sigma y}{n} = \frac{793}{22} = 36.04$

$n = 22$

$\Sigma x = 1221$, $\Sigma y = 793$

$\Sigma x^2 = 68857$

$\Sigma y^2 = 30125$

$\Sigma xy = 43708$

$\bar{x} = 55.5$

$\bar{y} = 36.04$

$\sigma_x = 7.04$

$\sigma_y = 8.36$

$S_x = 7.209$

$S_y = 8.5661$

$A = 51.47$

$B = -0.278$

$r = -0.234$

(iii)

$$S_x = \sqrt{\frac{\Sigma x^2}{n-1} - \frac{n\bar{x}^2}{n-1}}$$

from calculator;

$\Sigma x^2 = 68857$

$\bar{x} = 55.5$

$n = 22$

$\therefore S_x = 7.209$

$$S_y = \sqrt{\frac{\Sigma y^2}{n-1} - \frac{n\bar{y}}{n-1}}$$

from calculator;

$\Sigma y^2 = 30125$

$\bar{y} = 86.04$

$n = 22$

$\therefore S_y = 8.566$

(iii) x is good as its average is greater ($\bar{x} > \bar{y}$)

(iv) We have,

$$CV_x = \frac{S_x}{\bar{x}} \times 100\% = \frac{7.209}{55.5} \times 100\%$$

$$= 12.98\%$$

$$CV_y = \frac{S_y}{\bar{y}} \times 100\% = \frac{8.566}{36.04} \times 100\% =$$

$$= 23.07\%$$

Since $C.V_x < C.V_y$, x is consistent.

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

From calculator,

$\Sigma xy = 43708$
$\Sigma x = 1221$
$\Sigma y = 793$
$\Sigma x^2 = 68857$
$\Sigma y^2 = 30125$

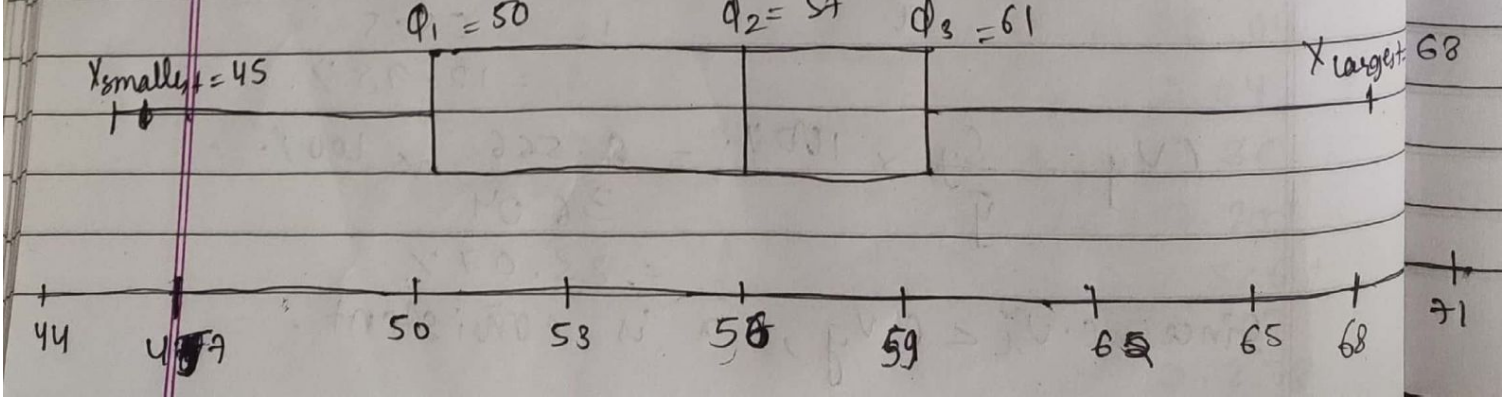
$\therefore r = -0.234$

(vi) $$SE(\bar{x} - \bar{y}) = \sqrt{\frac{\sigma x^2}{n_1} + \frac{\sigma y^2}{n_2}}$$

$= 2.33$

(vii)

$x_{smallest} = 45$
$Q_1 = 50$
$Q_2 = 57$
$Q_3 = 61$
$x_{largest} = 68$



(VIII)

incomplete

Remember:

Common critical value

	Confidence level	Critical value
$\alpha = 0.1$	90%	1.645
$\alpha = 0.05$	95%	1.960
$\alpha = 0.01$	0.99	2.575