DAYANANDA SAGAR COLLEGE OF ENGINEERING
(An Autonomous Institute Affiliated to VTU, Belagavi)
Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560078
DEPARTMENT OF MATHEMATICS

Course Material

COURSE	**MATHEMATICS FOR COMPUTER ENGINEERS**
COURSE CODE	**21MAT31A**
MODULE	**3**
MODULE NAME	**STATISTICS**
STAFF INCHARGE	**Dr. D. R. SASI REKHA**

Contents of the Module

Statistics

- ***Analyzing a data***
- ***Mean, Median, Mode***
- ***Standard deviation- Combination of two groups***
- ***Correlation, Covariance & Linear Regression***

STATISTICS

Introduction

Statistics is a science of facts and figures which may be readily available or obtained through the process of direct enquiry or enumeration. It deals with the methods of collecting, classifying and analyzing the data so as to draw some valid conclusions.

Basic Definitions and Formulae in Statistics

Classification

The significance of a large mass of statistical data known as raw data cannot be understood unless it is arranged in some definite manner. The process of arrangement in different groups is called classification.

Frequency table

It is a tabular arrangement consisting of various classes of uniform size known as class intervals and the number in each class known as frequency.

The difference between two consecutive vertical entries of the classes is known as the width of the class usually denoted by h. The average of the left and right end points of the class interval is known as the midpoint of the class usually denoted by x.

Mean, Variance and Standard Deviation

(i) If $x_1, x_2, \ldots\ldots\ldots x_n$ be a set of 'n' values of a variate x, then Mean, denoted by $\bar{x}$, variance denoted by V and Standard Deviation denoted by σ is defined as follows:

- Mean = $\bar{x} = \frac{\sum x}{n}$
- Variance = $V = \frac{\sum (x-\bar{x})^2}{n}$ (or) $V = \frac{\sum x^2}{n} - \bar{x}^2$
- Standard Deviation = $\sigma = \sqrt{V}$

(ii) For the grouped data in the form of a frequency distribution

- Mean = $\bar{x} = \frac{\sum fx}{\sum f}$
- Variance = $V = \frac{\sum f(x-\bar{x})^2}{\sum f}$ (or) $V = \frac{\sum x^2 f}{\sum f} - \bar{x}^2$
- Standard Deviation = $\sigma = \sqrt{V}$

Standard Deviation of the Combination of two groups

If m_1 , σ_1 be the mean and standard deviation of a sample of size n_1 and m_2 , σ_2 be the mean and standard deviation of a sample of size n_2 then the standard deviation σ of the combined ample of size $n_1 + n_2$ is given by

$$(n_1 + n_2)\sigma^2 = n_1\,\sigma_1^2 + n_2\,\sigma_2^2 + n_1\,D_1^2 + n_2\,D_2^2$$

where $D_i = m_i - M$; m being the mean of combined sample.

Median

(i) If the values of a variable are arranged in the ascending or descending order of magnitude:

- The median is the middle item if the number is odd
- The median is the mean of the two middle item if the number is even.

Thus the median is equal to the mid-value (ie) the value which divides the total frequency into equal parts

(ii) For the grouped data:

$$\boldsymbol{Median} = \boldsymbol{L} + \frac{\boldsymbol{H}}{\boldsymbol{F}}\left(\frac{\boldsymbol{N}}{\boldsymbol{2}} - \boldsymbol{C}\right)$$

where L = lower limit of the median
H = width of the median class
F = frequency of the median class
N = total frequency
C = cumulative frequency up to the class preceding the median class

Mode

(i) The mode is defined as that value of the variable which occurs most frequently (ie) the value of the maximum frequency.

(ii) For the grouped distribution it is given by the formula for regular distribution

$$\boldsymbol{Mode} = \boldsymbol{L} + \frac{\boldsymbol{H}\,\boldsymbol{f_1}}{\boldsymbol{f_1} + \boldsymbol{f_2}}$$

where L = lower limit of the mode

f_1 = excess of modal frequency over frequency of preceding class

f_2 = excess of modal frequency over following class

H = width of the model class

PROBLEMS:

1. Find the (i) mean, (ii) median, (iii) mode and (iv) standard deviation of a set of observations: 6, 8, 7, 5, 4, 9, 3, 3, 3

Solution:

(i) Mean = $\bar{x} = \frac{\sum x}{n} = \frac{6+\ 8+\ 7+\ 5+\ 4+\ 9+\ 3+3+3}{9} = 5.333$

(ii) set of observations in ascending order: 3 , 3 , 3 , 4 , 5 , 6 , 7 , 8 , 9

Median = middle term = 5

(iii) Mode = the number which occurs frequently = 3

(iv) Standard Deviation = $\sigma = \sqrt{V}$

$\therefore$ Variance = $V = \frac{\sum x^2}{n} - \bar{x}^2$

$$V = \frac{1}{9}\{6^2 + 8^2 + 7^2 + 5^2 + 4^2 + 9^2 + 3^2 + 3^2 + 3^2\} - 5.333^2$$

$$V = 4.6667$$

Standard Deviation = $\sigma = \sqrt{V} = \sqrt{4.6667} = 2.1603$

2. Find the (i) mean, (ii) median, (iii) mode and (iv) standard deviation for the following grouped data

Mid value	5	10	15	20	25	30	35
Frequency	2	5	8	10	7	4	1

Solution:

x	f	xf	x^2f	C.F
5	2	10	50	2
10	5	50	500	7
15	8	120	1800	15
20	10	200	4000	25
25	7	175	4375	32
30	4	120	3600	36
35	1	35	1225	37
	$\sum f$ = 37	$\sum fx$ = 710	$\sum x^2f$ = 15550	

(i) Mean = $\bar{x} = \frac{\sum fx}{\sum f} = \frac{710}{37} = 19.1892$

(ii) Here $N = \frac{37}{2} = 18.5$, which lies at x = 30, hence Median is 30

(iii) The value of x corresponding to the maximum frequency size 10 is 30. Hence mode is 30

(iv) Standard Deviation = $\sigma = \sqrt{V}$

$\therefore$ Variance = $V = \frac{\sum x^2 f}{\sum f} - \bar{x}^2 = \frac{15550}{37} - 19.1892^2 = 52.0453$

Standard Deviation = $\sigma = \sqrt{V} = \sqrt{52.0453} = 7.2142$

3. Find the (i) mean, (ii) median, (iii) mode and (iv) standard deviation for the following grouped data

Class	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
Frequency	3	16	26	31	16	8

Solution:

Class	f	x	fx	x^2f	C.F
0 - 10	3	5	15	75	3
10 - 20	16	15	240	3600	19
20 - 30	26	25	650	16250	45
30 - 40	31	35	1085	37975	76
40 - 50	16	45	720	32400	92
50 - 60	8	55	440	24200	100
	$\sum f$ = 100		$\sum fx$ =3150	$\sum x^2f$ = 114500	

(i) Mean = $\bar{x} = \frac{\sum fx}{\sum f} = \frac{3150}{100} = 31.5$

(ii) Here $\frac{N}{2} = \frac{100}{2} = 50$, which falls in the interval 30 – 40.

Hence $L = 30,\ H = 10,\ F = 31,\ C = 45$

$$Median\ =\ L + \frac{H}{F}\left(\frac{N}{2} - C\right) =\ 30 + \frac{10}{31}(50 - 45) = 31.6129$$

(iii) The maximum frequency size 31 which falls in the interval 30 – 40

Hence $L = 30,\ H = 10,\ f_1 = 31 - 26 = 5,\ f_2 = 31 - 16 = 15$

$$Mode\ =\ L + \frac{H\, f_1}{f_1 + f_2} = 30 + \frac{10(5)}{5 + 15} = 32.$$

(iv) Standard Deviation = $\sigma = \sqrt{V}$

$$\therefore\ \text{Variance} = V = \frac{\sum x^2 f}{\sum f} - \bar{x}^2 = \frac{114500}{100} - 31.5^2 = 152.75$$

Standard Deviation = $\sigma = \sqrt{V} = \sqrt{152.75} = 12.3592$

4. The scores obtained by two bats men X and Y in 10 matches are given below. Calculate mean and standard deviation and coefficient of variation for each batsman. Who is the better score getter and who is more consistent?

X	**30**	**44**	**66**	**62**	**60**	**34**	**80**	**46**	**20**	**38**
Y	**34**	**46**	**701**	**38**	**55**	**48**	**60**	**34**	**45**	**30**

Solution:

X	Y	X^2	Y^2
30	34	900	1156
44	46	1936	2116
66	70	4356	4900
62	38	3844	1444
60	55	3600	3025
34	48	1156	2304
80	60	6400	3600
46	34	2116	1156
20	45	400	2025
38	30	1444	900
$\sum X$ = 480	$\sum Y$ = 460	$\sum X^2$ = 26152	$\sum Y^2$ = 22626

Mean of X = $\bar{X} = \frac{\sum X}{n} = \frac{480}{10} = 48$

Mean of Y = $\bar{Y} = \frac{\sum Y}{n} = \frac{460}{10} = 46$

Variance of X = $\sigma_1^2 = \frac{\sum X^2}{n} - \bar{X}^2 = \frac{26152}{10} - 48^2 = 311.2$

Variance of Y = $\sigma_2^2 = \frac{\sum Y^2}{n} - \bar{Y}^2 = \frac{22626}{10} - 46^2 = 146.6$

Standard Deviation of X = $\sigma_1 = \sqrt{V} = \sqrt{311.2} = 17.6409$

Standard Deviation of Y = $\sigma_2 = \sqrt{V} = \sqrt{146.6} = 12.1078$

Coefficient of variation of X = $\frac{\sigma_1}{\bar{X}} * 100 = \frac{17.6409}{48} * 100 = 36.7519$

Coefficient of variation of Y = $\frac{\sigma_2}{\bar{Y}} * 100 = \frac{12.1078}{46} * 100 = 326.3213$

Since mean of X is greater than mean of Y, therefore X is the better score getter. And coefficient of variation of X is less than the coefficient of variation of Y, then X is more consistent than Y.

5. An analysis of monthly wages paid to the workers of two company A and B belonging to the same industry gives the following results:

Company	Number of workers	Mean monthly wages	Variance
X	500	186	81
Y	600	175	100

(i) Which company has a larger wages bill?

(ii) In which company is there greater variability in individual wages

(iii) Calculate the mean and standard deviation of wages of all the workers in the company A and B taken together.

Solution:

No. of worker in company X = $n_1 = 500$

Mean of X = $\bar{X} = \frac{\sum X}{n_1} = 186$ ➔ $\sum X = n_1\bar{X} = 500(186) = 93000$

No. of worker in company Y = $n_2 = 600$

Mean of Y = $\bar{Y} = \frac{\sum Y}{n_2} = 175$ ➔ $\sum Y = n_2\bar{Y} = 600(175) = 105000$

Variance of X = $\sigma_1^2 = 81$

Variance of Y = $\sigma_2^2 = 100$

Standard Deviation of X = $\sigma_1 = \sqrt{81} = 9$

Standard Deviation of Y = $\sigma_2 = \sqrt{100} = 10$

Coefficient of variation of X = $\frac{\sigma_1}{\bar{X}} * 100 = \frac{9}{186} * 100 = 4.84$

Coefficient of variation of Y = $\frac{\sigma_2}{\bar{Y}} * 100 = \frac{10}{175} * 100 = 5.71$

(i) $\sum X$ is lesser than $\sum Y$, therefore Y has larger wages bill.

(ii) Coefficient of variation of X is less than the coefficient of variation of Y, then Y has greater variability in individual wages.

(iii) Combined mean = $\frac{n_1\,\bar{X}_1+n_2\,\bar{Y}_2}{n_1+n_2} = \frac{500(186)+600(175)}{500+600} = 180$

Combined S.D:

$$\sigma^2 = \frac{1}{n_1+n_2}(n_1\,\sigma_1^2 + n_2\,\sigma_2^2 + n_1\,D_1^2 + n_2\,D_2^2)$$

$$\sigma^2 = \frac{1}{500+600}(500(81) + 600(100) + 500(186-180)^2 + 600(175-180)^2)$$

$$\sigma^2 = 121.3636$$

$$\therefore \quad \sigma = \sqrt{121.36} = 11.0165$$

Correlation, Covariance and Regression

Correlation and correlation coefficient

Co-variation of two independent magnitudes is known as correlation. If two variables x and y are related in such a way that increase or decrease in one of them corresponds to increase or decrease in the other. We say that the variables are positively correlated. Also if increase or decrease in one of them corresponds to decrease or increase in the other, the variables are said to be negatively correlated.

The numerical measure of correlation between x and y is known as ***Pearson's coefficient of correlation*** usually denoted by r and is refined as follows

$$r = \frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})(\boldsymbol{y}-\overline{\boldsymbol{y}})}{n\sigma_x\sigma_y}$$

Alternative formula for r

1. $r = \frac{\sum \boldsymbol{XY}}{\sqrt{\sum \mathbf{X^2}}\sqrt{\sum \mathbf{Y^2}}}$ where $\boldsymbol{X = x - \overline{x}}$; $\boldsymbol{Y = y - \overline{y}}$; $\sigma_x^2 = \frac{\sum(x-\overline{x})^2}{\boldsymbol{n}}$; $\sigma_y^2 = \frac{\sum(y-\overline{y})^2}{\boldsymbol{n}}$
2. $r = \frac{\sigma_x^2+\sigma_y^2-\sigma_z^2}{2\sigma_x\sigma_y}$ where $z = x - y$

Covariance:

Let the corresponding values of two variable X and Y given in ordered pairs then the covariance between X and Y is denoted by $Cov(X,Y)$ and it is defined as

$$\boldsymbol{Cov(X,Y)} = \frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})(\boldsymbol{y}-\overline{\boldsymbol{y}})}{\boldsymbol{n}}$$

Alternative formula for $Cov(X,Y)$

1. $Cov(X,Y) = E(XY) - E(X)\,E(Y)$; **where** $E(XY), E(X), E(Y)$ **are the corresponding mean**
2. $Cov(X,Y) = r\,\sigma_x\sigma_y$

Properties

- The coefficient of correlation numerically does not exceed unity
- If X and Y are independent then $Cov(X,Y) = 0$

Note

If $r = \pm 1$, we say that x and y are perfectly correlated and if $r = 0$, we say that x and y are non-correlated

Regression

Regression is an estimation of one independent variable in terms of the other. If x and y are correlated, the best fitting straight line in the least square sense gives reasonably a good relation between x and y.

The best fitting straight line of the form $y = ax + b$ (x being the independent variable) is called the regression line of y on x and $x = ay + b$ (y being the independent variable) is called the regression line of x on y

Equation of the regression lines

1. The regression line of y on x is $\boldsymbol{y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})}$
2. The regression line of x on y is $\boldsymbol{x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})}$

Alternative formula for regression lines

1. The regression line of y on x is $Y=\frac{\sum XY}{\sum X^2}(X)$
2. The regression line of x on y is $X=\frac{\sum XY}{\sum Y^2}(Y)$ where $\boldsymbol{X=x-\bar{x}\,;\ Y=y-\bar{y}}$

Note

1. $\boldsymbol{r\frac{\sigma_y}{\sigma_x}}$ and $\boldsymbol{r\frac{\sigma_x}{\sigma_y}}$ are known as the regression coefficients.
2. The product of the regression coefficients is equal to r^2. (ie) the coefficient of correlation by the lines of regression is $r=\pm\sqrt{(coeff.of\ x)(coeff.of\ y)}$
3. The angle between the lines of regression is $tan\theta=\frac{\sigma_x\,\sigma_y}{\sigma_x^2+\sigma_y^2}\left(\frac{1-r^2}{r}\right)$
4. If $r=\pm1$, $tan\theta=0$. Therefore $\theta=0$ which implies that the two regression lines coincide and hence the variables are perfectly correlated. Also if $r=0$, $tan\theta=\infty$. Therefore $\theta=\frac{\pi}{2}$ which implies that the lines are perpendicular and hence the variable are uncorrelated

PROBLEMS:

6. Compute (i) the coefficient of correlation, (ii) Covariance and (iii) the equation of the lines of regression for the data

x	1	2	3	4	5	6	7
y	9	8	10	12	11	13	14

Solution:

(i) We have $r = \frac{\sigma_x^2+\sigma_y^2-\sigma_z^2}{2\sigma_x\sigma_y}$ where $z = x - y$

x	y	z = x - y	x^2	y^2	z^2
1	9	-8	1	81	64
2	8	-6	4	64	36
3	10	-7	9	100	49
4	12	-8	16	144	64
5	11	-6	25	121	36
6	13	-7	36	169	49
7	14	-7	49	196	49
$\sum x$ = 28	$\sum y$ = 77	$\sum z$ = -49	$\sum x^2$ = 140	$\sum y^2$ = 875	$\sum z^2$ = 347

$$\bar{x} = \frac{\sum x}{n} = \frac{28}{7} = 4 \quad ; \quad \sigma_x^2 = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{140}{7} - 4^2 = 4$$

$$\bar{y} = \frac{\sum y}{n} = \frac{77}{7} = 11 \quad ; \quad \sigma_y^2 = \frac{\sum y^2}{n} - \bar{y}^2 = \frac{875}{7} - 11^2 = 4$$

$$\bar{z} = \frac{\sum z}{n} = \frac{-49}{7} = -7 \quad ; \quad \sigma_z^2 = \frac{\sum z^2}{n} - \bar{z}^2 = \frac{347}{7} - (-7)^2 = 0.57$$

Therefore $r = \frac{4+4-0.57}{2(2)(2)} = \mathbf{0.93}$

(ii) $Cov(X,Y) = r\,\sigma_x\sigma_y$

$= (0.93)(2)(2) = 3.72$

(iii) The line of regression are given by

The regression line of y on x is $\boldsymbol{y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})}$

$$\boldsymbol{y-11=(0.93)\frac{2}{2}(x-4)}$$

$$\boldsymbol{y-11=(0.93)(x-4)}$$

$$\boldsymbol{y=0.93x+7.28}$$

The regression line of x on y is $\boldsymbol{x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})}$

$$\boldsymbol{x-4=(0.93)\frac{2}{2}(y-11)}$$

$$\boldsymbol{x-4=(0.93)(y-11)}$$

$$\boldsymbol{x=0.93y-6.23}$$

7. Obtain (i) the equation of the lines of regression, (ii) the coefficient of correlation and (iii) Covariance for the following data

x	1	3	4	2	5	8	9	10	13	15
y	8	6	10	8	12	16	16	10	32	32

Solution:

(i) The line of regression are given by

The regression line of y on x is $Y=\frac{\sum XY}{\sum X^2}(X)$

The regression line of x on y is $X=\frac{\sum XY}{\sum Y^2}(Y)$ where $\boldsymbol{X=x-\bar{x}\,;\ Y=y-\bar{y}}$

$$\bar{x}=\frac{\sum x}{n}=\frac{70}{10}=7 \quad ; \quad \bar{y}=\frac{\sum y}{n}=\frac{150}{10}=15$$

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
1	8	-6	-7	36	49	42
3	6	-4	-9	16	81	36
4	10	-3	-5	9	25	15
2	8	-5	-7	25	49	35
5	12	-2	-3	4	9	6
8	16	1	1	1	1	1
9	16	2	1	4	1	2
10	10	3	-5	9	25	-15
13	32	6	17	36	289	102
15	32	8	17	64	289	136
$\sum x$ = 70	$\sum y$ = 150			$\sum X^2$ = 204	$\sum Y^2$ = 818	$\sum XY$ = 360

The regression line of y on x is $Y = \frac{360}{204}(X)$

$$y - 15 = 1.76(x - 7)$$

$$y = 1.76x + 2.68$$

The regression line of x on y is $X = \frac{360}{818}(Y)$

$$x - 7 = 0.44(y - 15)$$

$$x = 0.44y + 0.4$$

(ii) Coefficient of correlation by the lines of regression is

$$r = \pm\sqrt{(coeff.of\ x)(coeff.of\ y)}$$

$$r = \sqrt{(1.76)(0.44)} = 0.88$$

(iii) $Cov(X, Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n} = \frac{360}{10} = 36$

8. $8x - 10y + 66 = 0$ **and** $40x - 18y = 214$ **are the two regression lines. Find the mean of x's , y's and the correlation coefficient. Find** σ_y if $\sigma_x = 3$

Solution:

We know that regression lines passes through $\bar{x}$ and $\bar{y}$.

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

On solving we get $\bar{x} = 13$ and $\bar{y} = 17$

We shall now rewrite the equation of the regression lines to find the regression coefficients

$10y = 8x + 66$ ➔ $y = 0.8x + 6.6$

$40x = 18y + 214$ ➔ $x = 0.45y + 5.35$

Therefore coefficient of x = 0.8 and coefficient of y = 0.45

$$r = \pm\sqrt{(coeff.of\ x)(coeff.of\ y)} = \sqrt{(0.8)(0.45)} = 0.6$$

Also given $\sigma_x = 3$,

We have $r\frac{\sigma_y}{\sigma_x} = 0.8$

$(0.6)\sigma_y = 2.4$ ➔ $\sigma_y = 4$

9. Calculate the covariance of the following pairs of observation of two variables: $(1,4), (2,2), (3,4), (4,8), (5,9), (6,12)$

Solution:

$$E(XY) = \frac{\sum XY}{n} = \frac{4+4+12+32+45+72}{6} = \frac{169}{6}$$

$$E(X) = \frac{\sum X}{n} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6}$$

$$E(Y) = \frac{\sum Y}{n} = \frac{4+2+4+8+9+12}{6} = \frac{39}{6}$$

$$Cov(X,Y) = E(XY) - E(X)\,E(Y)$$

$$= \frac{169}{6} - \left(\frac{21}{6}\right)\left(\frac{39}{6}\right)$$

$$= 5.4167$$

10. Given r = 0.8, write down the equation of the lines of regression and hence find the most probable value of y when x =90. Given

	x-series	y-series
Mean	**18**	**100**
SD	**14**	**20**

Solution:

Given $\bar{x} = 18$; $\bar{y} = 100$; $\sigma_x = 14$; $\sigma_y = 20$; $r = 0.8$

The line of regression are given by

The regression line of y on x is $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

$$y - 100 = (0.8)\frac{20}{14}(x - 18)$$

$$y - 100 = (1.14)(x - 18)$$

$$y = 1.14x + 79.48 \text{ ----------(1)}$$

The regression line of x on y is $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

$$x - 18 = (0.8)\frac{14}{20}(y - 100)$$

$$x - 18 = (0.56)(y - 100)$$

$$x = 0.56y - 38$$

The most probable value of y when x = 70 is

(1) ➔ $y = 1.14x + 79.48 = 1.14(70) + 79.48 = 159.28$

11. If the coefficient of correlation between two variables x and y is 0.5 and the acute angle between their lines of regression is $\tan^{-1}\left(\frac{3}{5}\right)$. Show that $\sigma_y = 2\sigma_x$ or $\sigma_x = 2\sigma_y$

Solution:

Given r =0.5 and $\theta = \tan^{-1}\left(\frac{3}{5}\right)$ **(ie)** $tan\theta = \frac{3}{5}$

w.k.t The angle between the lines of regression is

$$tan\theta = \frac{\sigma_x\,\sigma_y}{\sigma_x^2 + \sigma_y^2}\left(\frac{1-r^2}{r}\right)$$

$$\frac{3}{5} = \frac{\sigma_x\,\sigma_y}{\sigma_x^2 + \sigma_y^2}\left(\frac{1-0.5^2}{0.5}\right)$$

$$\frac{3}{5} = \frac{\sigma_x\,\sigma_y}{\sigma_x^2 + \sigma_y^2}\left(\frac{3}{2}\right)$$

$$\frac{1}{5} = \frac{\sigma_x\,\sigma_y}{\sigma_x^2 + \sigma_y^2}\left(\frac{1}{2}\right)$$

$$5\sigma_x\,\sigma_y = 2(\sigma_x^2 + \sigma_y^2)$$

$$5\sigma_x\,\sigma_y = 2\sigma_x^2 + 2\sigma_y^2$$

$$2\sigma_x^2 + 2\sigma_y^2 - 5\sigma_x\,\sigma_y = 0$$

On factorizing we obtain $(2\sigma_x - \sigma_y)(\sigma_x - 2\sigma_y) = 0$

Thus we get $\boldsymbol{\sigma_y = 2\sigma_x}$ **or** $\boldsymbol{\sigma_x = 2\sigma_y}$

Problems for practice

1. Find the mean of the following by

Number	8	10	15	20
Frequency	5	8	8	4

2. The following is the frequency distribution of a random sample of weekly earnings of the employees. Calculate the average weekly

Weekly earning	10	12	14	16	18	20	22	24	26	28	30	32
No. of employees	3	6	10	15	24	42	75	90	79	55	36	26

3. Find the mean of the following

Class	0 – 10	10 – 20	20 – 30	30 – 40	40 - 50
Frequency	7	8	20	10	5

4. Find the mean of the following

Class	0 – 8	8 – 16	16 – 24	24 – 32	32 - 40	40 - 48
Frequency	8	7	16	24	15	7

5. The total sale (in thousands) of a particular item in a shop, on 10 consecutive days is reported by a clerk as 35, 29.6, 38, 30, 40, 41, 42, 45, 3.6, and 3.8. Calculate the average. Later it was found that there was a number 10 in the machine and the reports of 4^{th} to 8^{th} day were 10 more than the true values and in the last 2 days he put a decimal in wrong place (for example 3.6 was really 36). Calculate the true mean value.

6. For the two frequency distributions given below the mean calculated from the first was 25.4 and that the second was 32.5 find the value of the x and y.

Class	10 – 20	20 – 30	30 – 40	40 - 50	50 – 60
Frequency - 1	20	15	10	x	y
Frequency - 2	4	8	4	$2x$	y

7. Find the median of the following data

Number	1	2	3	4	5	6	7	8	9
Frequency	8	10	11	16	20	25	15	9	6

8. Find the median of the following data

Number	5	10	15	20	25	30	35	40	45
Frequency	29	224	465	582	634	644	650	653	655

9. Find the median of the following

Class	20 – 30	30 – 40	40 - 50	50 – 60	60 - 70
Frequency	3	5	20	10	5

10. A number of particular articles has been classified according to their weight. After drying for two week the same articles have again be weighted and similarly classified. It is known that the median weight in the first weight was 20.8302 while in the second weighting it was 17.3502. Some frequencies a and b in the first weighting and x and y in the second are missing. It is given that $a = \frac{x}{3}$ and $b = \frac{y}{2}$. Find out the missing frequencies.

Class	0 – 5	5 – 10	10 – 15	15 - 20	20 – 25	25 - 30
Frequency - 1	a	b	11	52	75	22
Frequency - 2	x	y	40	50	30	28

11. In a factory employing 3000 person, 5% earn less than Rs. 3 per hour, 580 earn from Rs. 3.01 to Rs. 4.5 per hour, 30% earn from Rs. 4.51 to Rs. 6 per hour. 500 earn from Rs. 6.01 to Rs. 7.5 per hour, 20% earn from Rs. 7.51 to 9 per hour and the rest earn Rs. 9.01 or more per hour. What is the median wage?

12. According to the census of 2021, the following are the population figures in thousands of 20 cities: 2000, 1180, 1785, 1500, 560, 782, 1200, 385, 1123, 222, 2001, 1178, 1780, 1550, 559, 780, 1250, 390, 1120, 225. Find the median.

13. Find the mode of the following data

Number	1	2	3	4	5	6	7	8
Frequency	4	9	16	25	22	15	7	3

14. The median and mode are given to be Rs. 25 and Rs. 24 respectively. Calculate the missing frequency.

Class	0 – 10	10 – 20	20 – 30	30 – 40	40 - 50
Frequency	14	x	27	y	15

15. Find the mode of the following distribution

Class	0 – 10	10 – 20	20 – 30	30 – 40	40 - 50	50 - 60	60 - 70
Frequency	5	8	7	12	28	20	10

16. The median and mode of the following wages are known to be Rs. 33.5 and Rs. 34 respectively. Find the value of x, y and z. Given total frequency is 230.

Class	0 – 10	10 – 20	20 – 30	30 – 40	40 - 50	50 - 60	60 - 70
Frequency	4	16	x	y	z	6	4

17. Calculate the mode form the following frequency distribution by the method of grouping

Number	4	5	6	7	8	9	10	11	12	13
Frequency	2	5	8	9	12	14	14	15	11	13

18. Calculate the standard deviation from the following frequency distribution

Number	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

19. For a group of 200 candidates, the mean and standard deviation of scores were found to be 40 and 15 respectively. Later on it was discovered the score 43 and 35 was misread as 34 and 53 respectively. Find the corrected standard deviation corresponding to the corrected figures.

20. Compute the standard deviation for the following data

Class interval	0-99	100-199	200-299	300-399	400-499	500-599	600-699	700-799
Frequency	10	54	184	264	246	40	1	1

21. The first group of the two samples has 100 items with mean 15 and standard deviation 3. If the whole group has 250 items with mean 15.6 and standard deviation $\sqrt{13.44}$. Find the standard deviation of the second group.

22. The number examined, the mean weight and standard deviation in each group of examination by three medical examination are given below. Find the mean weight and standard deviation of the entire data when grouped together

Medical examination	Number examined	Mean weight (lbs)	Standard deviation (lbs)
A	50	113	6
B	60	120	7

23. Calculate the covariance of the following pairs of observation of two variables: $(10,35), (15,20), (20,30), (25,30), (30,35), (35,38), (40,42), (45,30), (50,40)$

24. Find the Covariance by using co-efficient of correlation between industrial production and export using the following data and comment on the result.

Production (in tons)	55	56	58	59	60	60	62
Exports(in tons)	35	38	38	39	44	43	45

25. Find the covariance for the data given below

x	98	87	90	85	95	75
y	15	12	10	10	16	7

26. Calculate the Covariance by using correlation co-efficient for the following heights in inches of fathers (x) and their sons (y).

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

27. Find the Covariance by using co-efficient of correlation between industrial production and export using the following data and comment on the result.

Production (in tons)	55	56	58	59	60	60	62
Exports(in tons)	35	38	38	39	44	43	45

28. Find the correlation co-efficient between x and y from the given data:

X	21	23	30	54	57	58	72	78	87	90
Y	60	71	72	83	110	84	100	92	113	135

29. Find the correlation co-efficient between x and y from the given data:

x	78	89	97	69	59	79	68	57
y	125	137	156	112	107	138	123	108

30. Obtain the regression lines of y on x and x on y and hence find the correlation coefficient for the following data:

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

31. Obtain the regression lines of y on x and x on y and hence find the correlation coefficient for the following data:

x	1	2	3	4	5
y	2	5	3	8	7

32. Establish the formula $r = \frac{{\sigma_x}^2 + {\sigma_y}^2 - {\sigma_{x-y}}^2}{2\,\sigma_x\,\sigma_y}$

33. The two regression equations of the variables x and y are $x = 19.13 - 0.87y$ and $y = 11.64 - 0.50x$. Find (i) mean of *x*'s, (ii) mean of *y*'s and (iii) the correlation coefficient between *x* and *y*

34. The following results were obtained from records of age(x) and blood pressure (y) of a group of 10 men, given $\Sigma\,(x - \bar{x})(y - \bar{y}) = 1220$. Find the appropriate regression equation and use it to estimate the blood pressure of a man whose age is 45

	x	y
Mean	53	142
Variance	130	165

35. In a partially destroyed laboratory record of correlation data, the following result only are a variable, variance of x is 9, regression equation y on x and x on y are $4x - 5y + 33 = 0$, $20x - 9y - 107 = 0$ respectively. Calculate the coefficient of correlation, $\bar{x}$, $\bar{y}$ and σ_y

36. If θ is the acute angle between the two regression lines relating the variables *x* and *y*, show that $\tan\theta = \frac{\sigma_x\sigma_y}{{\sigma_x}^2+{\sigma_y}^2}\left(\frac{1-r^2}{r}\right)$

37. Find the co-efficient of correlation between x and y given $2\sigma_x = \sigma_y$ and the angle between the lines of regression is $\tan^{-1}\left(\frac{3}{5}\right)$

****