

**DEPARTMENT OF STATISTICS
FACULTY OF PHYSICAL SCIENCES
AHMADU BELLO UNIVERSITY, ZARIA**

STAT102: Lectures note

1.0 PRESENTATION OF DATA

Data collected has to be presented clearly before it can be analyzed affectively.

Statistical data can be presented as:

- **Raw Data:** These are collected data that have not been organized numerically.
- **Array:** An arrangement of data in ascending or descending order of magnitude is called an array. The difference between smallest and biggest number is called range.
- **Class:** A class is a group of items that are alike or simply a range of scores.
- **Frequency Data:** When we have large data the frequency distribution or frequency tabular arrangement of data by classes with the corresponding frequency has to be made.
- **Class Interval and Limit:** Class interval is any symbols defined as 5 – 10, 11 – 15 The ends value is called class limit.
- **Class Boundaries:** This is a number obtained by falling averages of the upper limit of a class and lower limit of the next higher class.
- **Size or Width of Class Interval:** The size of a class, denoted by C, is the number of element contained in a class. It is the difference between lower and upper class boundaries.
- **Class Mark:** It is the midpoint of a class.
- **Graphical Presentation of Data:** Data can be presented by mean of graphs. Some of the methods used are bar chart, histogram, ogive or cumulative curve, pie chart and frequency polygon.

2.0 MEASURE OF CENTRAL TENDENCY OR DESCRIPTIVE MEASURES

The following are the five measures of central tendency that are in common use:

- (i) Arithmetic mean or simple mean
- (ii) Median
- (iii) Mode
- (iv) Geometric mean
- (v) Harmonic mean

2.1 Arithmetic Mean

Arithmetic mean of a set of observation is their sum divided by the number of observations, e g , the arithmetic mean $\bar{x}$ of n observations, $x_1, x_2, \dots, x_n$ is given by:

$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n}\sum_{i=1}^n x_i$ in case of the frequency distribution $\frac{x_i}{f_i}, i = 1, 2, \dots, n$ where f_i is the frequency of the variable x_i ,

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^n f_ix_i}{\sum_{i=1}^n f_i} = \frac{1}{N} \sum_{i=1}^n f_ix_i$$

Where $\sum_{i=1}^n f_i = N$

In case of grouped or continuous frequency distribution, x is taken as the mid value of the corresponding class.

Example 1: Find the arithmetic mean of the following data:

5, 6, 7, 8, 9 and 10

Soln:

$$\text{Arithmetic mean} = \frac{5+6+7+8+9+10}{6} = \frac{45}{6} = 7.5$$

Example 2: Find the arithmetic mean of the following frequency distribution:

x: 1 2 3 4 5 6 7

f : 5 9 12 17 14 10 6

Solution

x	f	fx
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
TOTAL	73	299

$$\therefore \bar{x} = \frac{1}{N} \sum fx$$

$$= \frac{299}{73} = 4.09$$

Example 3: Calculate the arithmetic mean of the marks from the following table:

Marks: 0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
12	18	27	20	17	6

Marks	No. of Students (f)	Mid-point	fx
0 – 10	12	5	60
10 – 20	18	15	270
20 – 30	27	25	675
30 – 40	20	35	700
40 – 50	17	45	765
50 – 60	6	55	330
TOTAL	100		2,800

$$\begin{aligned}
 \text{Arithmetic mean } (\bar{x}) &= \frac{1}{N} \sum f x \\
 &= \frac{1}{100} \times 2,800 \\
 &= 28
 \end{aligned}$$

It may be noted that if the values fx or f large, the calculation of mean by formula handle. The arithmetic is reduced to a great extent by taking the deviations of the given values from any arbitrary point 'A' as explained below:

$$\text{Let } d_i = x_i - A$$

$$\text{Then } f_i d_i = f_i (x_i - A) = f_i x_i - A f_i$$

Summing both sides over i from 1 to n , we get

$$\sum_{i=1}^n f_i d_i = \sum_{i=1}^n f_i x_i - A \sum_{i=1}^n f_i = \sum_{i=1}^n f_i x_i - A \cdot N$$

$$\Rightarrow \frac{1}{N} \sum_{i=1}^n f_i d_i = \frac{1}{N} \sum_{i=1}^n f_i x_i - A = \bar{x} - A \text{ by adding 'A' to the both sides we have:}$$

$$\frac{1}{N} \sum_{i=1}^n f_i d_i + A = \frac{1}{N} \sum_{i=1}^n f_i x_i = \bar{x} \text{ where } \bar{x} \text{ is the arithmetic mean of the distribution.}$$

$$\therefore \bar{x} = A + \frac{1}{N} \sum_{i=1}^n f_i d_i.$$

This formula is much more convenient to apply than formula (**).

2.2 Median

Median of a distribution is the value of the variable which divided it into two equal parts. Let's the value such that the number of observations above it's equal to the number of observations bellow.

In case of ungrouped data, if the number of observations is odd then median is the middle value after the value have been arranged in ascending or descending order of magnitude. In case of even number of observations there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms. E. g

The median of 2, 1, 4, 3, 5, 5 e.g. 1, 2, 3, 4, 5 is 3

The median of 2, 3, 1, 5, 4, 6 e.g. 1, 2, 3, 4, 5, 6 is $\frac{3+4}{2} = \frac{7}{2} = 3.5$

In case of discrete frequency distribution median is obtained by considering the cumulative frequencies. The steps for calculating median are given below:

- (i) Find $\frac{1}{2}N$, where $N = \sum f_i$
- (ii) See the cumulative frequency (cf) just greater than $\frac{1}{2}N$.
- (iii) The corresponding value of x is median.

Example 4: Obtain the median for the following frequency distribution:

x :	1	2	3	4	5	6	7	8	9
f :	8	10	11	16	20	25	15	9	6

Solution

Computation of Median

x	f	cf
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120

TOTAL $N = 120$

$$\text{Here } N = 120 \Rightarrow \frac{N}{2} = 60$$

The cumulative frequency (cf) just greater than $\frac{1}{2}N$ is 65 and the value of x corresponding to 65 is 5. Therefore, median is 5.

2.2.1 Median for Continuous Frequency Distribution

In the case of continuous frequency distribution, the class corresponding to the cf just greater than $\frac{1}{2}N$ is called the median class and the value of median is obtained by the following formula.

$$\text{Median} = L + \frac{\left[\frac{N}{2} - \sum f_b\right]}{f_m} \times C$$

Where L is the lower limit of the median class, $\sum f_b$ is the sum of frequency of all the classes before the median class, f_m is the frequency of the median class, N is the total frequency and C is the class size of median class.

Example: The table below shows the distribution of weight of 30 students. Find the median of the following distribution

Weight (kg)	48 – 50	51 – 53	54 – 56	57 – 59	60 – 62	63 – 65
No. of students.	3	6	9	5	5	2

Solution

Class interval	x	f	cf
48 – 50	49	3	3
51 – 53	52	6	9
54 – 56	55	9	18
57 – 59	58	5	23
60 – 62	61	5	28
63 – 65	64	2	30

$$\text{Median} = L + \frac{\left[\frac{N}{2} - \sum f_b\right]}{f_m} \times C$$

$$L = 53.5 \quad \frac{N}{2} = 15, \sum f_b = 9$$

$$\begin{aligned}\text{Median} &= 54 + \frac{\left[\frac{30}{2} - 9\right]}{9} \times 3 \\ &= 53.5^+\end{aligned}$$

2.3 Mode

Mode is the value which occurs most frequently in a set of observations. In other words, mode is the value of the variable which is predominant in the series. Example, the mode of 2, 5, 6, 2, 4 and 3 is 2, because it appeared more than one unlike the other numbers which appear ones. In case of discrete frequency distribution, mode is the value of x corresponding to maximum frequency. For example, in the following frequency distribution:

$x:$	1	2	3	4	5	6	7	8
$f:$	4	9	16	25	22	15	7	3

Value of x corresponding to the maximum frequency 25 is 4. Hence mode is 4.

2.3.1 Mode for Continuous Frequency Distribution

In case of continuous frequency distribution, mode is given by the formula.

$$\text{Mode} = L + \frac{D_1}{D_1 + D_2} \times C$$

Where **L** is the lower limit of model class

D₁ is the difference between model class frequency and the frequency of the next upper class

D₂ is the difference between model class frequency and the frequency of the next lower class

C is the class size

Example 6: Find the mode for the following distribution:

Class-interval	0 – 10	10 – 20	20 – 30	30 – 40	40 - 50	50 – 60	60 – 70	70-80
Frequency	5	8	7	12	28	20	10	10

Solution

Class interval	Frequency
0 – 10	5
10 – 20	8
20 – 30	7
30 – 40	12
40 – 50	28
50 – 60	20
60 – 70	10
70 – 80	10

$$\text{Mode} = L + \left(\frac{D_1}{D_1 + D_2} \right) \times C$$

$$L = 40 \quad D_1 = 16 \quad D_2 = 8 \quad C = 10$$

$$\text{Mode} = 40 + \left(\frac{16}{16+8} \right) \times 10$$

$$= 40 + 0.667 \times 10$$

$$= 40 + 6.67$$

$$\text{Mode} = 46.67 \text{ (approx.)}$$

2.4 Geometric Mean

Geometric mean of a set of n observations is the n^{th} root of their product. Thus the geometric mean G , of n observations $x_i, i = 1, 2, \dots, n$ is given by:

$$G = (x_1 \cdot x_2 \dots x_n)^{1/n}$$

The computation is facilitated by use of logarithms. Taking logarithm of both sides.

$$\begin{aligned} \log G &= \frac{1}{n} (\log x_1 + \log x_2 + \dots \log x_n) \\ &= \frac{1}{n} \sum_{i=1}^n \log x_i \end{aligned}$$

$$\therefore G = \text{Antilog} \left(\frac{1}{n} \sum_{i=1}^n \log x_i \right).$$

In case of frequency distribution $x_i | f_i (i = 1, 2, \dots, n)$ geometric mean, G is :

$$G = (x_1^{f_1} \cdot x_2^{f_2} \dots x_n^{f_n})^{1/N}, \text{ where } N = \sum_{i=1}^n f_i$$

taking logarithm of both sides, we get

$$\begin{aligned} \log G &= \frac{1}{N} (f_1 \log x_1 + f_2 \log x_2 + \dots + f_n \log x_n) \\ &= \frac{1}{N} \sum_{i=1}^n f_i \log x_i \end{aligned}$$

Thus we find that logarithm of geometric mean is the arithmetic mean of the logarithms of the given value, we get:

$$G = \text{Antilog} \left(\frac{1}{N} \sum_{i=1}^n f_i \log x_i \right).$$

Example 7: The geometric mean of 3, 9 and 27 is $= (3 \cdot 9 \cdot 27)^{1/3} = 9$

Example 8: compute the geometric mean for the following data:

10, 110, 120, 50, 52, 80, 37, 60

Solution

Let us prepare the following table:

Size of item (x)	$\log x$	Value of $\log x$
10	$\log 10$	1.0000
110	$\log 110$	2.0414
120	$\log 120$	2.0792
50	$\log 50$	1.6990
52	$\log 52$	1.7160
80	$\log 80$	1.9031
37	$\log 37$	1.5682
60	$\log 60$	1.7782
$n = 8$		$\Sigma \log x = 13.7851$

$$\text{Log } G = \frac{1}{n} \sum \log x$$

$$= \frac{13.7851}{8}$$

$$\text{Log } G = 1.723$$

$$G = \text{antilog } (1.723)$$

$$= 52.84$$

Example 9: Compute the geometric mean for the ff. data:

$x:$	12	13	14	15	16	17
$f:$	5	4	4	3	2	1

Solution

x	$\log x$	f	$fx \log x$
12	1.0792	5	5.3960
13	1.1139	4	4.4556
14	1.1461	4	4.5844
15	1.1761	3	3.5283
16	1.2041	2	2.4092
17	1.2304	1	1.2304
$n = \Sigma f = 19$			$\Sigma \log x = 13.7851$

$$\begin{aligned}
 \text{Log} &= \frac{\Sigma f \log x}{n} \\
 &= \frac{21.6029}{19} = 1.137 \\
 \therefore G &= \text{Antilog}(1.137) \\
 &= 13.71
 \end{aligned}$$

2.5 Harmonic Mean

Definition:

The harmonic mean of n items $x_1, x_2, x_3, \dots, x_n$ is defined as harmonic mean.

$$\text{H.M} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}}$$

For example, the harmonic mean of 2, 4 and 5 is

$$\frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{5}} = \frac{3}{\frac{10 + 5 + 4}{20}} = \frac{3}{\frac{19}{20}} = 3 \div \frac{19}{20} = 3 \times \frac{20}{19}$$

$$= \frac{60}{19} = 3.16$$

Harmonic mean of frequency distribution data. Let $x_1, x_2, x_3, \dots, x_n$ be n items which occur with frequency $f_1, f_2, f_3, \dots, f_n$ respectively. Then their Harmonic mean is given by:

$$\text{Harmonic mean} = \frac{f_1 + f_2 + f_3 + \dots + f_n}{\left(\frac{f_1}{x_1} + \frac{f_2}{x_2} + \frac{f_3}{x_3} + \dots + \frac{f_n}{x_n}\right)} = \frac{\sum f_i}{\sum f_i \times \frac{1}{x_i}}$$

Example 10: Compute the harmonic mean for the following data:

Marks obtained	20	21	22	23	24	25
No. of Students	4	2	7	1	3	1

Solution

Let us compute the following table:

x	$\frac{1}{x}$	f	$f \times \frac{1}{x}$
20	0.05000	4	0.2000
21	0.04762	2	0.09524
22	0.04545	7	0.31815
23	0.04348	1	0.04348
24	0.04167	3	0.12501
25	0.04000	1	0.04000
	TOTAL	18	$\sum x^{1/x} = 0.82188$

$$\text{Now } \sum f x \frac{1}{x} = 0.82188$$

$$n = 18$$

$$\therefore \text{Harmonic mean} = \frac{18}{0.82188}$$

$$= 21.9$$

Example 11: The interest paid on each of three different sums of money yielding 3%, 4% and 5% simple interest respectively is the same. What is the average yield percent on the total sum interested?

Soln:

$$H.M = \frac{3}{\frac{1}{3} + \frac{1}{4} + \frac{1}{5}} = \frac{3}{\frac{20 + 15 + 12}{60}} = \frac{3 \times 60}{47} = \frac{180}{47} = 3.83\%$$

3.0 MEASURE OF DISPERSION

The meaning of dispersion is scatteredness. It helps in finding out the variability of the data or scatteredness of individual items in a given distribution. In other words, the degree to which numerical data tend to spread about an average value is called the dispersion of the data.

The absolute measures can be divided into the following four positional measures:

1. Range
2. Quartile Deviation or Semi-inter Quartile Deviation
3. Mean Deviation
4. Standard Deviation

3.1 Range

It is the simplest measure of dispersion. It is the difference between the minimum and maximum items of series. It can also be called absolute range.

$$\text{Range} = x_{\max} - x_{\min} = L - S$$

Where L is the large value and S is the smallest value of the term.

Example 12: In a series 20, 21, 22, 25, 30, 32, 37, 47, 65, the range is $65 - 20 = 45$.

$$\begin{aligned}\text{Coefficient of range or Relative range} &= \frac{\text{absolute range}}{\text{sum of the two extremes}} \\ &= \frac{L-S}{L+S}\end{aligned}$$

Referring to example 12 the coefficient of range is obtain as:

$$= \frac{L-S}{L+S} = \frac{65-20}{65+20} = \frac{45}{85} = \frac{9}{17}$$

3.2 Quartile Deviation or Semi-Inter Quartile Range

Quartile deviation is a measure of dispersion based on the upper quartile (Q_3) and Lower quartile (Q_1) of a series. It is half of the difference between the Upper quartile and the Lower quartile.

$$\text{QUARTILE DEVIATION} = \frac{Q_3 - Q_1}{2}$$

Where Q_3 = Third or Upper quartile

Q_1 = First or Lower quartile

$$\text{Coefficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example 13: if the first quartile is 104 and quartile deviation is 18, find the third quartile.

Soln:

$Q_1 = 104$ $Q_3 = ?$ Quartile deviation = 18

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

$$\therefore = \frac{Q_3 - 104}{2}$$

$$\Rightarrow 18 \times 2 = Q_3 - 104$$

$$36 + 104 = Q_3$$

$$\Rightarrow 140 = Q_3$$

$\therefore$ Third quartile: $Q_3 = 140$

Example 14: Calculate the quartile deviation and coefficient of quartile deviation from the following data:

Age in years	20	30	40	50	60	70	80
No. of members	3	61	132	153	140	51	3

Solution

Age in years (x)	No. of members (f)	Cumulative freq. (cf)
20	3	3
30	61	64
40	132	196
50	153	349
60	140	489
70	51	540
80	3	543

Here $N = 543$

$$Q_1 = \text{value of } \left(\frac{N+1}{4}\right)^{th} \text{ item} = \left(\frac{543+1}{4}\right) = 136^{th} \text{ item} = 40 \text{ years.}$$

$$Q_3 = \text{value of } 3\left(\frac{N+1}{4}\right)^{th} \text{ item} = 136 = 408^{th} = 60 \text{ years.}$$

$$\therefore \text{Quartile Deviation} = \frac{Q_3 - Q_1}{2} = \frac{60 - 40}{2} = 10 \text{ years and coefficient of}$$

$$Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{60 - 40}{60 + 40} = \frac{20}{100} = 0.2$$

Example 15: The marks in Arithmetic of a group of students are given below. Find quartile deviation and coefficient of quartile deviation.

Class-interval	No. of students	Class-interval	No. of students
Below 25	7	50 – 55	13
25 – 30	12	55 – 60	24
30 – 35	15	60 – 65	13
35 – 40	18	65 – 70	13
40 – 45	25	70 – 75	21
45 – 50	33	75 – 80	23

Solution: Computation of Quartile Deviation

Class-interval	No. of students (f)	cf
Below 25	7	7
25 – 30	12	19
30 – 35	15	34
35 – 40	18	52
40 – 45	25	77
45 – 50	33	110
50 – 55	13	123
55 – 60	24	147
60 – 65	13	160
65 – 70	13	173
70 – 75	21	194
75 – 80	23	217

$$N = \sum f = 217$$

$$Q_1 = \text{size of } \left(\frac{N}{4}\right)^{th} \text{ item} = \text{value of } \left(\frac{217}{4}\right) = 54.25^{th} \text{ item}$$

$\Rightarrow Q_1$ lies in the interval 40 – 45.

$$Q_1 = L + \left[\frac{\frac{N}{4} - \sum f_b}{f_q} \right] \times C$$

$$= 40 + \left[\frac{\frac{217}{4} - 52}{25} \right] \times 5$$

$$= 40 + \left(\frac{54.25 - 52}{25} \right) \times 5$$

$$\therefore Q_1 = 40 + 0.45 = 40.45\%$$

Also $Q_3 = \text{value of } \frac{3N}{4} \text{ th item} = \text{value of } \left(3 \times \frac{217}{4}\right) = 162.75^{th} \text{ item}$ which lies in the interval 65 – 70

$$Q_3 = L + \left[\frac{\frac{3N}{4} - \sum f_b}{q_3} \right] \times C$$

Where L is the lower class limit of the Q_3 class, N is the total frequency, $\sum f_b$ is the sum of frequency before frequency of the Q_3 class f_{q3} is the frequency of the Q_3 class and C is the class size or width.

$$Q_3 = 65 + \left[\frac{162.75 - 160}{13} \right] \times 5$$

$$= 65 + \frac{13.75}{13} = 65 + 1.06 = 66.06\%$$

$$\therefore \text{Quartile deviation} = \frac{Q_3 - Q_1}{2} = \frac{66.06 - 40.45}{2} = \frac{25.61}{2} = 12.81$$

$$\begin{aligned} \text{And coefficient of quartile deviation} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{66.06 - 40.45}{66.06 + 40.45} \\ &= \frac{25.61}{106.51} = 0.24 \end{aligned}$$

3.3 Mean Deviation or Average Deviation

Mean deviation of a set of observations of a series is the arithmetic mean of all the deviation, without their algebraic signs, taken from its central value (mean or medium or mode).

Now add up all n observation to get $\sum_{i=1}^n |x_i - \bar{x}|$

Mean Deviation: $D_x = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$, where $\bar{x}$ is the arithmetic mean, the vertical lines stand for positive value.

Similarly, mean deviation from the median Md for ungrouped data is denoted by

$MD_{\text{median}} = \frac{1}{n} \sum |x_i - Md|$ and mean deviation from the mode z for ungrouped

data $MD_r = \frac{1}{n} \sum |x_i - Z|$

Example 16: The marks obtained by 10 students in an examination were as follows:

70, 65, 68, 70, 75, 73, 80, 70, 83, 86

Find the mode, median and mean deviation about the mean.

Solution:

Arranging the data in the ascending order we have 65, 68, 70, 70, 70, 73, 75, 80, 83, 86.

$\Rightarrow$ mode = 70 ($\because$ 70 occurs many times than any number in the series e.g. 3 times)

$\Rightarrow$ median = Average of 5th and 6th item = $\frac{70+73}{2} = 71.5$

$$\begin{aligned}\Rightarrow \text{mean} &= \frac{\sum x}{n} = \frac{65+68+70+70+70+73+75+80+83+86}{10} \\ &= \frac{740}{10} = 74\end{aligned}$$

$$\therefore \text{mean deviation about mean} = \frac{\sum |x - \bar{x}|}{n}$$

$$= \frac{|65 - 74| + |68 - 74| + |70 - 74| + |70 - 74| + |70 - 74| + |73 - 74| + |75 - 74| + |80 - 74| + |83 - 74| + |86 - 74|}{10}$$

$$\begin{aligned}&= \frac{9 + 6 + 4 + 4 + 4 + 1 + 1 + 6 + 9 + 12}{10} \\ &= \frac{56}{10} = 5.6\end{aligned}$$

Please try or compute the:

- (i) mean deviation about the mode
- (ii) mean deviation about the median

3.3.1 Mean Deviation for Grouped Data

Let $x_1, x_2, x_3, \dots, x_n$ occur with frequencies $f_1, f_2, f_3 \dots f_n$ respectively and let $\sum f = n$ and M can be either mean or median or mode, then mean deviation is given by the formula.

$$\text{Mean deviation} = \frac{\sum f|x-M|}{\sum f}$$

Example 17: Find the mean and mean deviation about the mean for the sample observation on the weight (gm's) of a certain product.

9, 12, 10, 11, 8, 13, 11, 12, 10, 11, 12, 11, 8, 11, 16

Solution:

x	f	fx	$ x - 11 $	$f x - 11 $
8	2	16	3	2
9	1	9	2	2
10	2	20	1	2
11	6	66	0	0
12	3	36	1	3
13	1	13	2	2
16	1	16	5	5
	16	176		20

$$\text{Mean} = \frac{\sum fx}{\sum f} = \frac{176}{16} = 11$$

$$\text{Mean deviation} = \frac{\sum f|x - \bar{x}|}{\sum f} = \frac{20}{16} = 1.25$$

3.4 Standard Deviation

Standard deviation is the most important and commonly used measure of dispersion. It measures the absolute dispersion or variability of a distribution. A small standard deviation means a high degree of uniformity in the observations as well as homogeneity in the observations as well as homogeneity of the series.

Definition: Standard deviation is the positive square root of the average of squared deviation taken from arithmetic mean.

Let x be random variate which takes on n values, $x_1, x_2, x_3, \dots, x_n$, then the standard deviation of these n observations is given by Standard Deviation:

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

Where $\bar{x} = \frac{\sum x}{n}$ is the mean of these observations.

ALTERNATIVELY:

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$$

Coefficient of Standard Deviation = $\frac{\sigma}{\bar{x}}$

Example 18: Find the standard deviation of 3, 4, 5, 6

Solution:

Here $n = 4, \sum X = 3 + 4 + 5 + 6 = 18$

$$\sum X^2 = 3^2 + 4^2 + 5^2 + 6^2 = 9 + 16 + 25 + 36 = 86$$

$$\begin{aligned} \therefore \sigma &= \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{86}{4} - \left(\frac{18}{4}\right)^2} \\ &= \sqrt{21.5 - (4.5)^2} = \sqrt{21.5 - 20.25} = \sqrt{1.25} \\ &= 1.12 \text{ nearly.} \end{aligned}$$

Or

$$\text{Mean} = \frac{3+4+5+6}{4} = \frac{18}{4} = 4.5$$

x	$(x - \bar{x})$	$(x - \bar{x})^2$
3	-1.5	2.25
4	-0.5	0.25
5	0.5	0.25
6	1.5	2.25
		$\sum (x - \bar{x})^2 = 5$

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \sqrt{\frac{5}{4}} = \sqrt{1.25}$$

$$= 1.12$$

3.4.1 Standard Deviation – Discrete Series or Grouped Data

The standard deviation for the discrete series or grouped data is given by the formula.

$$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}}$$

OR

$$\sigma = \sqrt{\frac{\sum fx^2}{\sum f}} = \left(\frac{\sum fx}{\sum f} \right)^2$$

Example 19: Find the standard deviation from the following data. And also find the coefficient of variation.

Size of the item(x)	10	12	13	14	15	16
Frequency (f)	2	7	11	15	10	1

Solution:

x	f	fx	fx^2	$(x - \bar{x})$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
10	2	20	200	-2.8	7.84	15.68
11	7	77	847	-1.8	3.24	22.68
12	11	132	1584	-0.8	0.64	7.04
13	15	195	2535	0.2	0.04	0.6
14	10	140	1960	1.2	1.44	14.4
15	4	60	900	2.2	4.84	19.36
16	1	16	256	3.2	10.24	10.24

$$\sum f = 50, \quad \sum fx = 640, \quad \sum f(x - \bar{x})^2 = 90, \quad \sum fx^2 = 8282$$

$$mean = \frac{\sum fx}{\sum f} = \frac{\sum fx}{n} = \frac{640}{50} = 12.8$$

$$\sigma = \sqrt{\frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n}\right)^2} = \sqrt{\frac{8282}{50} - \left(\frac{640}{50}\right)^2}$$

$$= \sqrt{165.64 - 163.84} = \sqrt{1.8} = 1.342$$

$$\text{OR } \sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{n}} = \sqrt{\frac{90}{50}} = \sqrt{1.8} = 1.342$$

4.0 CORRELATION ANALYSIS

Correlation is a statistical tool which studies the relationship between two variables, and correlation analysis involves various methods and technique used for studying and measuring the extent of the relationship between two variables.

Definition: The relationship between two variables such that a change in one variable results in corresponding greater or smaller change in the other variable is known as correlation. The coefficient of correlation between the two variable x, y is generally denoted by r or r_{xy} or $P(x,y)$.

Covariance: Before we study the correlation analysis we introduce the concept of covariance between two quantitative variables X and Y . Let the corresponding values of the two variables X and Y on the given set of n units of observation be given by the ordered pairs. $(x_1, y_1) (x_2, y_2), (x_3, y_3), \dots (x_n, y_n)$.

Then the covariance between X and Y is denoted by $cov(X,Y)$. It is defined as:

$$COV(x, y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

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

Where $\bar{x}$ and $\bar{y}$ are the means of X and Y series respectively, i.e.

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}, \bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

Example: calculate the covariance of the following pairs of observations of the two variants X and Y .

(1,6) (2,9) (3,6) (4,7) (5,8) (7,12) (8,3) (9,17) (10,1)

Solution:

$$\sum x_i = 1+2+3+4+5+6+7+8+9+10 = 55$$

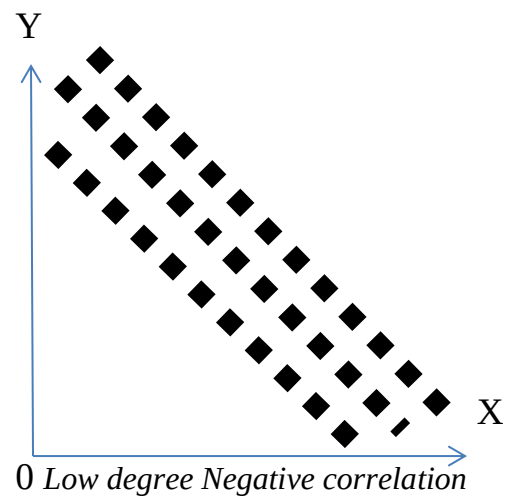
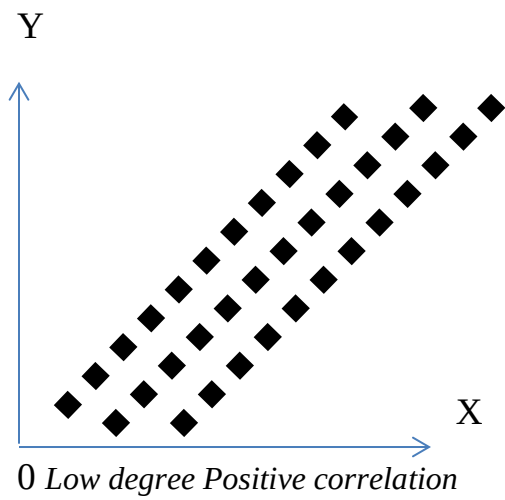
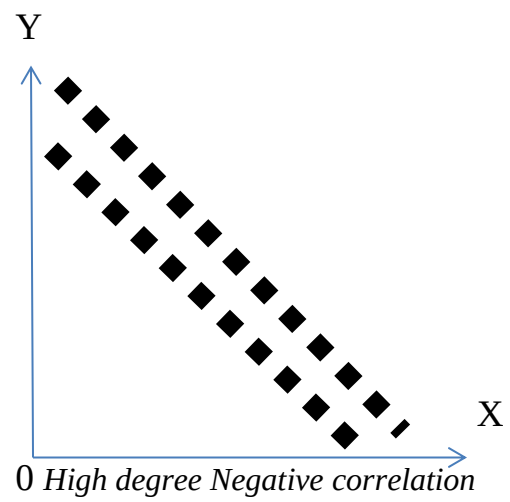
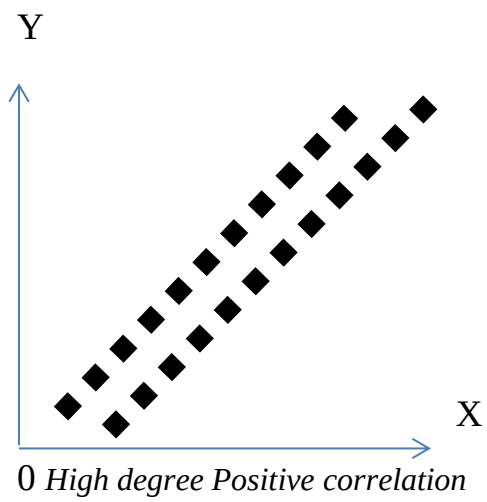
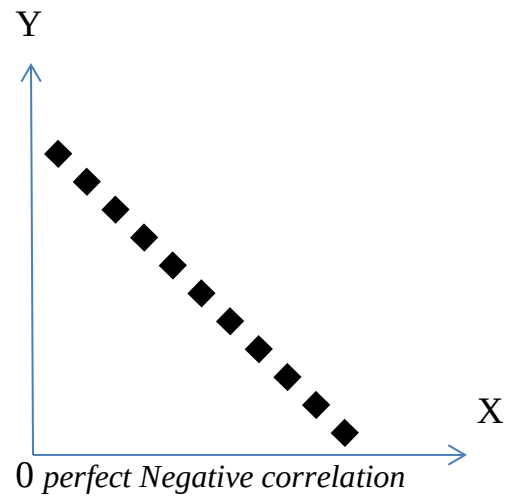
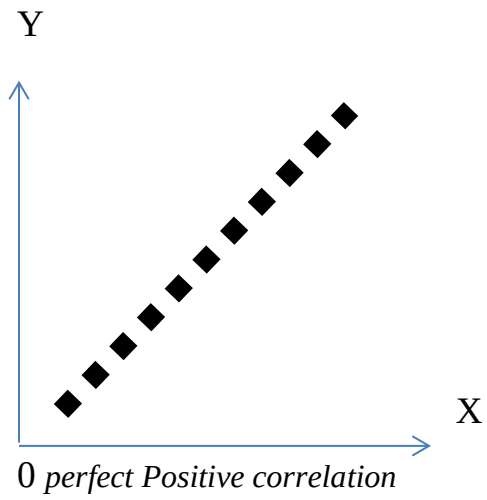
$$\sum y_i = 6+9+6+7+8+5+12+3+17+1 = 74$$

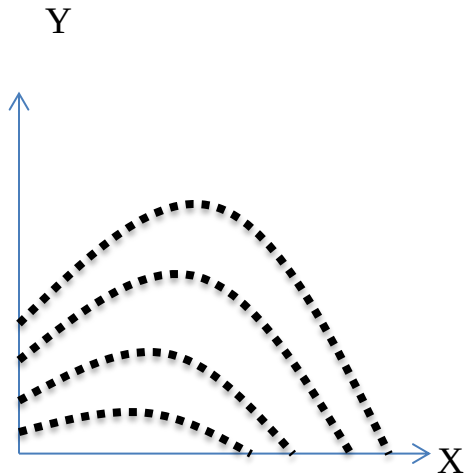
$$\begin{aligned}\sum x_i y_i &= (1 \times 6) + (2 \times 9) + (3 \times 6) + (4 \times 7) + (5 \times 8) + (6 \times 5) + (7 \times 12) + (8 \times 3) + (9 \times 17) + (10 \times 1) \\ &= 6+18+18+28+40+30+84+24+153+10 = \mathbf{411}\end{aligned}$$

$$\begin{aligned}\therefore COV(X_i Y) &= \frac{1}{n} \left(\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i \right) \\ &= \frac{1}{10} \left[411 - \frac{1}{10} \times 55 \times 74 \right] = \frac{1}{10} (411 - 407) = 0.4\end{aligned}$$

4.1 Types of Correlation

- (i) Positive Correlation
- (ii) Negative or Inverse correlation
- (iii) Linear Correlation
- (iv) Perfectly Linear Correlation
- (v) Perfect Correlation
- (vi) Direct or perfect Positive Correlation
- (vii) Inverse or Perfect Negative Correlation
- (viii) High degree Positive Correlation
- (ix) High degree Negative Correlation
- (x) Low degree Positive or Negative Correlation
- (xi) No Correlation





0 No Correlation

4.2 Methods to Study Correlation

The intensity of correlation between two variates x and y can be measured by the following methods:

- (i) Scatter or Dot diagram method
- (ii) Karl Pearson's Coefficient of correlation
- (iii) Spearman's Rank correlation method
- (iv) Two-way frequency table method
- (v) Concurrent Deviation method.

4.2.1 Karl Pearson's Coefficient of Correlation

$$P = (X, Y) = \frac{COV.(X, Y)}{\sqrt{COV(X) \times COV(Y)}}$$

$$P = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\{\sum_{i=1}^n (x_i - \bar{x})^2\} \{\sum_{i=1}^n (y_i - \bar{y})^2\}}}$$

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

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

Example: Calculate the Coefficient of correlation for the following data:

(1,2), (2,4), (3,8), (4,7), (5,10), (6,5), (7,14), (8,16), (9,2), (10,20)

Solution: here n = 10

$$\sum x_i = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 = 55$$

$$\sum y_i = 2 + 4 + 8 + 7 + 10 + 5 + 14 + 16 + 2 + 20 = 88$$

$$\sum x_i^2 = 1 + 4 + 16 + 25 + 36 + 49 + 64 + 81 + 100$$

$$\sum y_i^2 = 4 + 16 + 64 + 49 + 100 + 25 + 196 + 256 + 40 + 400 = 1114$$

$$\sum x_i y_i = 2 + 8 + 24 + 28 + 50 + 30 + 98 + 128 + 18 + 200 = 586$$

$$P(X, Y) = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right) \left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

$$= \frac{586 - \frac{55 \times 88}{10}}{\sqrt{\left(385 - \frac{55 \times 55}{10}\right) \left(1114 - \frac{88 \times 88}{10}\right)}}$$

$$= \frac{586 - 484}{\sqrt{(385 - 302.5)(1114 - 7745)}} = \frac{102}{\sqrt{2808.75}} = \frac{102}{167.35} = 0.61$$

4.3 Spearman's Rank Correlation Coefficient

The coefficient of rank correlation is based on the various values of the variates and is denoted by R. It is applied to the problems in which data cannot be measured quantitatively but qualitative assessment is possible such as beauty, honesty etc. in this case the best individual is given rank number 1, next rank 2 and so on. The coefficient of rank correlation is given by the formula:

$$R = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

Where D^2 is the square of the difference of corresponding ranks, and n is the number of pairs of observations.

Example: Two judges in a beauty competition rank the 12 entries as follows:

X :	1	2	3	4	5	6	7	8	9	10	11	12
Y:	12	9	6	10	3	5	4	7	8	2	11	1

Calculate the rank correlation coefficient between X and Y.

Solution: $n = 12$

Rank X	Rank Y	D	D^2
1	12	-11	121
2	9	-7	49
3	6	-3	9
4	10	-6	36
5	3	2	4
6	5	1	1
7	4	3	9
8	7	1	1
9	8	1	1
10	2	8	64
11	11	0	0
12	1	11	11

$$\sum D^2 = 416$$

$$R = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6 \times 416}{12(12^2 - 1)}$$

$$= 1 - \frac{6 \times 416}{1728 - 12} = 1 - \frac{2496}{1716} = 1 - 1.4545$$

$$= -0.4545$$

Example: The marks obtained by the students in physics and mathematics are as follows:

Physics(X): 35 23 47 17 10 43 9 6 28

Maths (Y): 30 33 45 23 8 49 12 4 31

Compute their ranks in the two subjects and the coefficient of correlation of ranks.

Soluton:

Assign rank1 to the highest value in both the series. We have the following table:

X	Y	$x_i = \text{Rank in } X$	$y_i = \text{Rank in } Y$	$D = x_i - y_i$	D^2
35	30	3	5	-2	4
23	33	5	3	2	4
47	45	1	2	-1	1
17	23	6	6	0	0
10	8	7	8	-1	1
43	49	2	1	1	1
9	12	8	7	1	1
6	4	9	9	0	0
28	31	4	4	0	0

$$\sum D^2 = 12$$

$$R = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6 \times 12}{9(81 - 1)} = 1 - \frac{72}{9 \times 80}$$

$$= 1 - \frac{1}{10} = \frac{9}{10} = 0.9$$

5.0 REGRESSION ANALYSIS

“Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data”.

There are two types of variables in regression analysis these are:

- (i) Independent variable
- (ii) Dependent variable

The variable whose value is influenced or is to be predicted is called Dependent variable, whereas the variable which influences the value or is used for prediction is called independent variable.

5.1 Types of Regression

- (i) Simple Regression: - The regression analysis confined to be study of only two variables at a time is called the simple regression.
- (ii) Multiple Regression: - The regression analysis for studying more than two variables at a time is known as multiple regression.

This course is confined only to the study of simple regression.

5.2 Lines of Regressions

1. Line of regression of X on Y. It is the line which gives the best estimate for the values of X for a specified value of Y. it is given: $X = a + bY$ where a is intercept of the line and b is the slope of the line X on Y.
2. Line of regression of Y on X. it is the line which gives the best estimate for the values of Y for any specified values of X. It is given by: $Y = a + bX$ where a is the intercept of the line and b is the slope of the line Y on X.

5.3 Method of Least Squares

The method of least squares is a mathematical device. It is used for obtaining the equation of a curve which fits best to a given set of observation.

5.3.1 Direct Method: -

Equation of a straight line is: $y = a + bx$. Normal equations for obtaining the values of the constants a and b are:

$$(i) \quad \Sigma y = Na + b \Sigma x \text{ and}$$

$$(ii) \quad \Sigma xy = \Sigma x + b \Sigma x^2.$$

The equation of the best fit is the straight line $y = a + bx$, where the values of a and b are obtained by solving the normal equations.

Example: Fit a straight line of Y on X from the following data.

X	0	1	2	3	4	5	6
Y	2	1	3	2	4	3	5

Solution:

The regression line of Y on X

$$Y = a + bx \text{ ----- (A)}$$

The two normal equations to solve a and b are

$$\Sigma y = Na + b \Sigma x \text{ and}$$

$$\Sigma xy = \Sigma x + b \Sigma x^2.$$

X	Y	X ²	XY	Y ²
0	2	0	0	4
1	1	1	1	1
2	3	4	6	9
3	2	9	6	4
4	4	16	16	16
5	3	25	15	9
6	5	36	30	25

$$\Sigma x = 21, \Sigma y = 20, \Sigma x^2 = 91, \Sigma xy = 74$$

Substituting the values of Σx , Σx^2 , Σxy and $N = 7$, we get

$$20 = 7a + 21b \text{ ----- (1)}$$

$$74 = 21a + 91b \text{ ----- (2)}$$

Solving equations (1) and (2) we get $a = 1.357$ and $b = 0.5$

Substituting the values of $a = 1.357$, and $b = 0.5$ in equation (A), we get $y = 1.357$,
+ $0.5x$ which is the line of best fit by least squares method.

$\therefore$ The regression line of x on y

$$\therefore x = a + by \text{ ----- (B)}$$

The two normal equations to evaluate a and b are:

$$\Sigma y = aN + b \Sigma xy$$

$$\Sigma xy = a \Sigma y + b \Sigma y^2.$$

After substituting we have:

$$21 = 7a + 20b \text{ ----- (3)}$$

$$74 = 20a + 68b \text{ ----- (4)}$$

Solving equation (3) and (4) we have $a = -0.6842$, $b = 1.2895$
Substituting the values of $a = -0.6842$ and $b = 1.2895$ in equation (B), we have
 $x = -0.6842 + 1.2895y$

6.0 MATHEMATICAL EXPECTATION

Mathematical expectation of a random variable is obtained by multiplying each probable value of the variable by its corresponding probability and then adding these products.

6.1 Laws of Expectation:

- (i) If **a** is any constant, then the expectation of **a** is equals to a i.e. $E(a) = a$.
- (ii) If **b** is any constant, then $E(bx) = b E(x)$
- (iii) If $x_1, x_2, x_3, \dots, x_n$ are n random variables then
 - a. $E(x_1 + x_2 + x_3 + \dots, x_n) = E(x_1) + E(x_2) + E(x_3) + \dots + E(x_n) = \sum_{i=1}^n E(x_i)$
 - b. $E(x_1, x_2, x_3, \dots, x_n) = E(x_1) E(x_2) E(x_3) \dots E(x_n)$ provided $x_1, x_2, x_3, \dots, x_n$ are stochastically independent

$$\prod_{i=1}^n x_i$$

Example (1)

The p.d.f. of a random variable x is given by

$$f(x) = \begin{cases} Cx^1 & \text{for } 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

Find the $E(x)$

Solution:

We first find the value of C for which $E(x)$ is a p.d.f

$$\int_0^3 x^3 dx = 1$$

$$C = \frac{4}{81}$$

Example (2)

A continuous random variable x has probability d.f. $f(x)$ defined by

$$f(x) = \begin{cases} 6x(1-x) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the mean and variance of x

Solution:

The mean

$$\begin{aligned} \mu &= E(X) \\ &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_0^1 6x^2(1-x)dx \\ &= \frac{1}{2} \end{aligned}$$

The variance

$$\begin{aligned} \mu &= E(x - \mu)^2 \\ &= E(x^2) - \mu^2 \\ E(X^2) &= \int_0^1 (6x^3 - 6x^4)dx \\ &= \frac{6}{20} \end{aligned}$$

$$\begin{aligned} \text{Var}(x) &= E(X^2) - \mu^2 \\ &= \frac{6}{20} - \frac{1}{4} \\ &= \frac{1}{20} \end{aligned}$$

7.0 PROBABILITY

Definition: If an experiment has n mutually-exclusive, equally and exhaustive cases, out of which m are favorable to the happening of the event A , then the probability of the happening of A is denoted by $P(A)$ and is defined as:

$$P(A) = \frac{m}{n} = \frac{\text{No. of cases favourable to } A}{\text{Total (Exhaustive) number of cases}}$$

7.1 Some Important Terms

Random Experiment: Random experiment means an action or an operation which can produce some well define outcomes.

Sample Space: A set of all possible outcomes from an experiment is called a sample space.

Event: A set of sub-collection of a number of sample points under a definite rule is called an event.

Exhaustive cases: It is the total number of all the possible outcomes of an experiment. For example, when we throw a dice, then any one of the six faces (1, 2, 3, 4, 5, 6) may turn up and therefore, there are six possible outcomes. Hence there are six exhaustive cases or events in throwing a dice.

Mutually exclusive Events: If in an experiment the occurrence of an event prevents or rules out the happening of all other events in the same experiment, then these events are said to be mutually exclusive events.

For example: in tossing coin, the events head and tail are mutually exclusive events, because if the outcome is head, then the possibility of getting a tail in the same trial is ruled out.

Equally Likely Events: Events are said to be equally likely if there is no reason to expect any one in preference to other.

Favorable Cases: The cases which ensure the occurrence of an event are said to be favorable to the event.

Independent and dependent events: Two or more events are said to be independent if the happening of any one does not depend on the happening of the other. Events which are not independent are called dependent events.

Example: What is the probability of getting an even number in a single throw with a dice?

Solution:

The possible cases in the throw of a dice are six i.e. 1, 2, 3, 4, 5, 6.

Favorable cases are those which are marked with 2, 4, 6 and these are three in numbers.

$$\therefore \text{Probability of getting an even number} = \frac{3}{6} = \frac{1}{2}$$

Example: What is the probability of getting tail in a throw of a coin?

Solution:

When we toss a coin, there are two possible outcomes, i.e. Head or tail. In this case the number of possible cases $n = 2$

No. of favorable cases = $m = 1$

$$\therefore \text{Probability of getting a tail} = \frac{m}{n} = \frac{1}{2}$$

Example: A bag contains 6 white balls, 9 black balls. What is the probability of drawing a black ball?

Solution:

The total number of equally likely and exhaustive cases = $n = 6 + 9 = 15$

Number of favourable cases = $m = 9$

$$\therefore \text{Probability of drawing a black ball} = \frac{9}{15} = \frac{3}{5}$$

Example: An unbiased coin is tossed three times find the probability of getting:

- (a) An exactly two head
- (b) At least two tail
- (c) Almost one head
- (d) Less than two head
- (e) At least one head
- (f) More than one head

Solution:

$$S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$$

$$(a) E = \{HHT, HTH, THH\} = 3 \quad n(s) = 8 \quad P(\text{exactly 2 H}) = \frac{3}{8}$$

$$(b) E = \{TTH, THT, HTT, TTT\} = 4 \quad P(\text{at least 2 T}) = \frac{4}{8} = \frac{1}{2}$$

$$(c) E = \{THT, TTH, HTT, TTT\}, \quad P(\text{almost 1 H}) = \frac{4}{8} = \frac{1}{2}$$

$$(d) P(\text{less than 2 H}) = \frac{3}{8}$$

$$(e) P(\text{at least 1 H}) = \frac{7}{8}$$

$$(f) P(\text{more than 1 H}) = \frac{4}{8} = \frac{1}{2}$$

Example: A bag contains three black, four red and five white balls. If a ball is picked at random from the bag. Find the probability that it's

- (a) A black
- (b) Not white
- (c) Red or black

Let B denote a black ball, R red ball, W white

$\therefore n(s) = 12$, since there are only $3B + 4R + 5W = 12$ balls

$$(a) P(B) = \frac{3}{12} = \frac{1}{4}$$

$$(b) P(\text{not white}) = P(W^c) = 1 - P(W), \text{ but } P(W) = \frac{5}{12}$$

$$\therefore (W^c) = 1 - \frac{5}{12} = \frac{7}{12}$$

$$(c) P(R \text{ or } B) = P(R \cup B) = P(R) + P(B) = \frac{4}{12} + \frac{3}{12} = \frac{7}{12}$$

Example: The probability that machine A will be performing a usual function in 5yrs time is $\frac{1}{4}$, while the probability that machine B will still be operating usefully at the end of the same period is $\frac{1}{3}$. Find the probability that in 5yrs time:

- (a) Both machines will be performing a usual function
- (b) Neither will be operating
- (c) Only machine B will be operating
- (d) At least one of the machine will be operating

Solution:

$$P(A) = \frac{1}{4}, P(B) = \frac{1}{3}, q(A) = \frac{3}{4}, q(B) = \frac{2}{3}$$

$$(a) P(A \cap B) = P(A) \times P(B) = \frac{1}{4} \times \frac{1}{3} = \frac{1}{12}$$

$$(b) q(A \cap B) = q(A) \times q(B) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}$$

$$(c) P(B \cap \bar{A}) = P(B) \times P(\bar{A}) = P(B) \times q(A) = \frac{1}{3} \times \frac{3}{4} = \frac{1}{4}$$

$$(d) 1 - q(A \cap B) = 1 - \frac{1}{2} = \frac{1}{2}$$

Or

$$(P(A) \times q(B)) + (q(A) \times P(B)) + P(A) \times P(B)$$

$$\left(\frac{1}{4} \times \frac{2}{3}\right) + \left(\frac{3}{4} \times \frac{1}{3}\right) + \left(\frac{1}{4} \times \frac{1}{3}\right)$$

$$\frac{2}{12} + \frac{3}{12} + \frac{1}{12} = \frac{6}{12} = \frac{1}{2}$$

Example: Two dice are thrown simultaneously, find the probability of getting:

- (a) A sum of 6
- (b) A sum divisible by 6
- (c) A sum of 6 or sum divisible by 6
- (d) A total of at least 10
- (e) A multiple of 2 on one die and multiple of 3 on the other die.

Solution :

Let E_1 denote event of sum of 6

E_2 denote event a sum divisible by 6

E_3 denote event a sum of 6 or sum divisible by 6

E_4 denote event a total of at least 10

E_5 denote event multiple of 2 on one die and multiple of 3 on the other die

there are 36 outcomes, $n(S) = 36$

$$(a) P\left(E_1\right) = \frac{n(E_1)}{n(S)} = \frac{5}{36}$$

$$(b) P(E_2) = \frac{n(E_2)}{n(S)} = \frac{6}{36}$$

$$(c) P(E_1 \text{ or } E_2) = P(E_3) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

$$\text{But } P(E_1) = \frac{5}{36}, P(E_2) = \frac{6}{36}, P(E_1 \cap E_2) = \frac{5}{36}$$

$$\therefore P(E_3) = \frac{5}{36} + \frac{6}{36} - \frac{5}{36} = \frac{6}{36} = \frac{1}{6}$$

$$(d) P(E_4) = \frac{6}{36} = \frac{1}{6}$$

$$(e) P(E_5) = \frac{11}{36}$$

Example: A card is drawn from a table of 52 packed card. Find the probability that:

- (a) A card is 3 of Heart
- (b) A card is 3 Spade
- (c) A card is 3 Jack
- (d) A black card
- (e) A king of black colour
- (f) An ace or club card
- (g) A king or red card
- (h) A card is spade or diamond

Solution:

(a) Number of favourable outcomes that a card is 3 of Heart is 1 out of 52

$$\therefore P(\text{a 3 of Heart}) = \frac{1}{52}$$

(b) A card of spade is 13 out of 52 cards

$$\therefore P(\text{a card of spade}) = \frac{13}{52} = \frac{1}{4}$$

(c) $P(\text{a card is jack}) = \frac{4}{52} = \frac{1}{13}$

(d) $P(\text{a black card}) = \frac{13}{52} + \frac{13}{52} = \frac{26}{52} = \frac{1}{2}$

(e) Number of king in spade card is 1 out of 52, number of king in spade card is

$$1 \text{ out of } 52, \therefore P(\text{king of black color}) = \frac{1}{52} + \frac{1}{52} = \frac{2}{52} = \frac{1}{26}$$

(f) $P(\text{a card is ace}) = \frac{4}{52}$, $P(\text{a card is club}) = \frac{13}{52}$, $P(A \cup C) = P(A) + P(C) -$

$$P(A \cap C) = \frac{4}{52} + \frac{13}{52} = \frac{17}{52} = \frac{17}{52}$$

(g) $P(\text{a king or red card}) = P(K \cup R) = P(k) + P(R) - P(k \cap R)$ But

$$P(k) = \frac{4}{52}, P(R) = \frac{26}{52}, P(k \cap R) = \frac{2}{52}, P(K \cup R) = \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52} = \frac{7}{13}$$

(h) $P(\text{card is spade or diamond}) = P(S) + P(D) = \frac{13}{52} + \frac{13}{52} = \frac{26}{52} = \frac{1}{2}$

8.0 BINOMIAL DISTRIBUTION

Let us assume that trials are independent and the probability of success is same in each trial. Let us claim that we have n trials, then the probability of happening of an event r times and failing $(n - r)$ times in any specified order is $P^r q^{n-r}$. But the total number of ways in which the event can happen r times exactly in n trials is $C(n,r)$. these $C(n,r)$ ways are equally likely, mutually exclusive and exhaustive.

Therefore, the probability for r successes and $(n - r)$ failures in n trials in any order is ${}^nC_r P^r q^{n-r}$. It can also be expressed in the form.

$P(x = r) = P(r) = {}^nC_r P^r q^{n-r}$; $r = 0, 1, 2, 3, \dots, n$; where $P(x = r)$ or $P(r)$ is the probability distribution of a random variable x of r successes. Giving different values to r , i.e putting $r = 0, 1, 2, 3, \dots, n$; we get the corresponding probabilities, ${}^nC_0 q^n, {}^nC_1 q^{n-1}, {}^nC_2 q^{n-2} P^2, {}^nC_3 q^{n-3} P^3 \dots, P^n$.

8.1 Mean and Variance of Binomial Distribution

The probability distribution of Binomial distribution for r successes in n trials is given by $P(r) = {}^nC_r q^{n-r} P^r$

It has mean = np

And variance = npq

Example: in a Binomial distribution, the mean and standard deviation are 12 and 2 respectively. Find n and p .

Solution:

We are given that: mean = $np = 12$ and S.D = $\sqrt{npq} = 2$

$$\Rightarrow npq = 4$$

$$\text{Now } = \frac{npq}{np} = \frac{4}{12} = \frac{1}{3} \rightarrow q = \frac{1}{3}$$

$$\therefore P = 1 - q = 1 - \frac{1}{3} = \frac{2}{3}$$

$$\text{Also } np = 12 \Rightarrow n \times \frac{2}{3} = 12 \Rightarrow n = 18$$

$$\text{Hence } n = 18, P = \frac{2}{3}$$

Example: Using the formula for Binomial distribution, find the probability of rolling at most 2 sixes in 5 rolls of a dice.

Solution:

Here $n = 5$

Probability of rolling a six on a dice = $P = \frac{1}{6}$

$$\therefore q = 1 - P = 1 - \frac{1}{6} = \frac{5}{6}$$

$$P(x \leq 2) = P(x = 0) + P(x = 1) + P(x = 2)$$

Hence our Binomial distribution formula are:

$${}^nC_r P^r q^{n-r}$$

$$\begin{aligned} & {}^5C_0 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^5 + {}^5C_1 \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^4 + {}^5C_2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 \\ &= \left(\frac{5}{6}\right)^5 + 5\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^4 + \frac{5 \times 4}{2 \times 1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 \\ &= \frac{5^4}{6^5} [5 + 5 + 2] = 12 \times \frac{5^4}{6^5} = \frac{625}{648} \end{aligned}$$

Example: if 10 coins are tossed 100 times, how many times would you expect 7 coins to fall head upward?

Solution:

The probability of getting a head in a single toss of a coin is $P = \frac{1}{2}$

$$\therefore q = 1 - P = 1 - \frac{1}{2} = \frac{1}{2}$$

Also we are given $n = 10$, $N = 100$ and $r = 7$

$\therefore$ The required frequency = $N \cdot {}^nC_r P^r q^{n-r}$

$$= 100 \times {}^{10}C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^3 = 100 \frac{10!}{7!3!} \left(\frac{1}{2}\right)^{10}$$

$$= 100 \frac{10 \times 9 \times 8}{3 \times 2 \times 1} \left(\frac{1}{2}\right)^{10} = \frac{375}{32} = 11.7 \cong 12$$

Example: Eight coins are thrown simultaneously. Find the probability of getting at least six heads.

Solution:

Let P denote the Probability of getting a head

Let q denote the probability of not getting a head

Then $P = \frac{1}{2}$, $q = \frac{1}{2}$.

$$e^\lambda, \frac{e^{-\lambda} \lambda}{1!} + \frac{e^{-\lambda} \lambda^2}{2!}, \dots, \frac{e^{-\lambda} \lambda^x}{x!}$$

Also the formula of Poisson distribution for computing the theoretical or relative frequencies of a random variable x is given by:

Theoretical frequency = $\left(\frac{e^{-\lambda} \lambda^x}{x!}\right)$, where N is number of trials, λ = mean.

The following are the statistical measures of the Poisson distribution.

1. Mean = $\lambda = np$
2. Variance = $\lambda = np$
3. Standard deviation = $\sqrt{\lambda} = \sqrt{np}$

Example: The mortality rate for a certain disease is 7 in 1000. What is the probability for just 2 deaths on account of this disease in a group of 400?

Give $e^{-2.8} = 0.06$

Solution:

Here $P = \frac{7}{1000}$ and $n = 400$

$$\lambda = np = 400 \times \frac{7}{1000} = 2.8$$

$$P(2) = \frac{\lambda^2 e^{-\lambda}}{2!} = \frac{(2.8)^2 e^{-2.8}}{2!} = \frac{2.8 \times 2.8}{2} \times 0.06 = 0.2352 =$$

23.52%

$\therefore$ Probability of getting at least six heads when 8 coins are thrown simultaneously:

$$= P(6) + P(7) + P(8)$$

$$= {}^8C_6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^2 + {}^8C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^1 + {}^8C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^0$$

$$= \left(\frac{1}{2}\right)^8 ({}^8C_6 + {}^8C_7 + {}^8C_8) = \frac{1}{256} [28 + 8 + 1] = \frac{37}{256}$$

9.0 POISSON DISTRIBUTION

Definition: The probability distribution of a random variable X is said to have a Poisson distribution if it takes only non – negative values and if it's distribution is given by:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \dots$$

Where λ is the parameter and $e = 2.7183$ or $e = 1 + \frac{1}{1} + \frac{1}{2!} + \frac{1}{3!} + \dots$

Hence the probability of 0, 1, 2, ... x successes are given by:

Example: It is given that 3% of the electric bulbs manufactured by a company are defective. Using Poisson distribution, find the probability that a sample of 100 bulbs will contain no defective bulb?

Solution:

Let P be the probability of defective bulb. Then $P = \frac{3}{100}$, $n = 100$

$$\therefore \lambda = np = 100 \times \frac{3}{100} = 3$$

We know that: $P(r) = \frac{e^{-\lambda} \lambda^r}{r!}$

Now probability that the sample will contain no defective bulb is

$$P(0) = \frac{e^{-3} (3)^0}{0!} = e^{-3} = 0.05$$

Example: The average number of a customers, who appear at a counter of a certain bank per minute is 2. Find the probability that during a given minute:

(a) No customer appears

(b) Three or more customers appears $P(X \geq 3) = 1 - P(X \leq 2)$