Business Statistics Notes For B.Com Part 1

Chapter 1

MEASURES OF CENTRAL TENDENCY

An important characteristic of data is given by a quantity known as the average of the data.

If two populations (or samples) are to be compared in respect of some characteristic, data are obtained for the populations (or samples) and then "averages" of the two data are compared to arrive at some conclusions. The word "Average" in statistical terminology is referred to as "Central Tendency" of the data.

> *A central Tendency is defined as a single value of the data which truly represents the whole data.*

There are different methods to compute the central tendency of any data known as "measures of central tendency". These measures are:

Arithmetical Measures:

1. Arithmetic Mean.
2. Geometric Mean.
3. Harmonic Mean.

Positional Measures :

4. Median
5. Mode

4.1 ARITHMETIC MEAN

Arithmetic Mean is the most commonly used measure of central tendency and due to this reason it is simply called Mean. It is defined as the sum of the values divided by the number of values in the raw data. Mean of a sample of n values, called the sample mean, denoted by $\bar{x}$ (Read as x bar), is

$$\bar{x} = \frac{\Sigma x}{n} \qquad \text{Sample Mean}$$

If the data is not a sample, rather the entire population of N values, the population mean denoted by μ is

$$\mu = \frac{\Sigma x}{N} \qquad \text{Population Mean}$$

Example: 4.1.1

A sample of personnel records shows that the number of sick leaves in days for each of five employees of an organization are 7, 4, 2, 7 and 5. Find the mean of these values.

Solution :

$$n = 5 \text{ values} \qquad \Sigma x = 25 \text{ days}$$

$$\bar{x} = \frac{\Sigma x}{n}$$

$$= \frac{25}{5}$$

$$\bar{x} = 5 \text{ days.}$$

4.2 WEIGHTED MEAN

The mean of a data gives equal importance (or weights) to each of the values of raw data. In some cases all values in the raw data do not have the same importance. A weighted mean is used to assign any degree of importance to each value of the data by choosing appropriate weights for these values.

The weighted mean of n values is

$$\bar{x}_w = \frac{\Sigma w.x}{\Sigma w} \qquad \text{Weighted Mean}$$

where w are the weights assigned to the values of data

Example: 4.2.1

If an investor buys 200 shares at a price of Rs.45/- each and 250 shares at Rs.36/- each, find the mean price per share.

Solution:

x = Price per share

w = No. of shares.

x	w	$w.x$
45 36	200 250	9000 9000
Total	450	18000

$$\bar{x}_w = \frac{\Sigma w.x}{\Sigma w} = \frac{18000}{450}$$

= Rs. 40/- per share

Example: 4.2.2

Marks obtained by a student in three tests of different durations are recorded below. Calculate the mean of the marks obtained by the student.

Test	Marks	Duration (Minutes)
1 2 3	30 80 70	30 120 30

Solution:

Without considering the duration of the tests, the mean is

$$\bar{x} = \frac{\Sigma x}{n} = \frac{180}{3} = 60 \text{ marks}$$

But this mean of 60 marks is not correct because every test has been considered equally important whereas they are not. Considering the duration of the tests as their relative importance (weights), the weighted mean would be computed as:

Marks x	Duration w	$w.x$
30 80 70	30 120 30	900 9600 2100
Total	180	12600

$$\bar{x}_w = \frac{\Sigma w.x}{\Sigma w} = \frac{12600}{180}$$

= 70 marks

Which is the correct mean in this case

$\bar{x} = \frac{126}{6} = 21$ units. $\bar{y} = \frac{3252}{6} =$ Rs. 542

Example: 4.3.3

There are 320 labourers in a large industrial complex, working in different wage grades. The average daily wage is calculated as Rs 73.00. A group of 80 more labourers has been appointed with average daily wage of Rs. 61.00

Find the average daily wage of all the labourers in the complex

Solution:

$n_1 = 320$ $\qquad n_2 = 80$

$\bar{x}_1 =$ Rs. 73 $\qquad \bar{x}_2 =$ Rs. 61

$n_1 . \bar{x}_1 =$ Rs. 23360 $\qquad n_2 . \bar{x}_2 =$ Rs. 4880

$$\bar{x}_c = \frac{n_1 . \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$= \frac{Rs.\ 23360 + Rs.\ 4880}{320 + 80}$$

$$= \frac{28240}{400}$$

$$= \text{Rs. } 70.60$$

4.4 ARITHMETIC MEAN OF GROUPED DATA

In a frequency distribution the formula $\bar{x} = \frac{\Sigma x}{n}$ cannot be directly applied because of the fact that the actual values of the data are lost. To overcome this difficulty an assumption is made without which the calculation of mean would have been impossible.

This assumption is:

The mid-points of the intervals are the values of data

Once this assumption is made, sum of the values of the data can be approximately obtained by taking the total of the product of mid-points and corresponding frequencies. This total is then divided by the total of frequencies to get the mean of the frequency distribution.

The formula to calculate the mean of a frequency distribution is thus:

$$\bar{x} = \frac{\Sigma f \cdot x}{\Sigma f}$$ Mean of a grouped data

Mean will be denoted by μ when the frequency distribution represents a population.

Example: 4.4.1

The raw data of example 2.2.1 about the number of cars sold are reproduced below:

7	8	13	10	9	10	5	12	8	6
10	11	12	5	10	11	10	5	9	13
8	12	8	8	10	15	7	6	8	8
5	6	9	7	14	8	13	5	5	14

This raw data are now converted into a frequency distribution with only 4 classes of size 3 each.

Classes	Frequency
5 — 7	12
8 — 10	17
11 — 13	8
14 — 16	3

Calculate the mean of the raw data and of the frequency distribution and compare the two results.

Solution :

Mean of ungrouped data of 40 values is given by

$$\bar{x} = \frac{\Sigma x}{n}$$

$$= \frac{360}{40}$$

$$= 9 \text{ cars.}$$

Mean of grouped data is calculated below:

Classes	f	x	f.x
5 — 7	12	6	72
8 — 10	17	9	153
11 — 13	8	12	96
14 — 16	3	15	45
Total	40	—	366

Different Grouping error of freq. dist

$$\bar{x} = \frac{\Sigma f.x}{\Sigma f} = \frac{366}{40}$$

or $\bar{x} = 9.15 \simeq 9$ cars

4.5 GEOMETRIC MEAN (G.M)

Geometric Mean is defined only for non-zero positive values. It is the nth root of the product of n values in the data. In symbols

$$G.M. = (\Pi x)^{1/n} = \sqrt[n]{\Pi x}$$

$$\text{Where } \Pi x = x_1 . x_2 . x_3 \ldots\ldots x_n$$

Example: 4.5.1

Calculate G.M. for the following values:

2, 5, 12, 18, 3, 8

Solution:

$$G.M. = (2 \times 5 \times 12 \times 18 \times 3 \times 8)^{1/6}$$

$$= (51840)^{1/6}$$. This can be evaluated using any scientific calculator.

$$= 6.1$$

Another way to calculate G.M. involves logarithm, that is

$$G.M. = \text{Antilog}\left[\frac{\Sigma \log x}{n}\right]$$

$$G.M. = \text{Antilog}\left[\frac{\log 2 + \log 5 + \log 12 + \log 18 + \log 3 + \log 8}{6}\right]$$

$$= \text{Antilog}\left[\frac{0.3010 + 0.6990 + 1.0792 + 1.2553 + 0.4771 + 0.9031}{6}\right]$$

$$= \text{Antilog}\left[\frac{4.7147}{6}\right]$$

$$= \text{Antilog } (0.7858)$$
$$= 6.1$$

The geometric mean is used mainly to find the average of ratios, rates of change, economic indices and the like. It is preferred when the data array follows a pattern of "geometric progression".

4.6 WEIGHTED GEOMETRIC MEAN

If weights (relative importance) are assigned to the values of the data, a weighted geometric mean is calculated using the formula

$$\text{G.M.} = \text{Antilog}\left[\frac{\Sigma w \cdot \log x}{\Sigma w}\right]$$

4.7 GEOMETRIC MEAN OF GROUPED DATA

The geometric mean of a frequency distribution is a special case of weighted geometric mean when frequencies are considered as weights.

$$\text{Therefore} \quad \text{G.M.} = \text{Antilog}\left[\frac{\Sigma f \cdot \log x}{\Sigma f}\right]$$

and hence log of G.M. is the arithmetic mean of log values of data.

Example: 4.7.1

Calculate Geometric Mean for the following frequency distribution.

C. I.	10 — 19	20 — 29	30 — 39	40 — 49	50 — 59	60 — 69
f	7	13	22	11	6	5

Solution :

C.I	f.	x	log x	f. log x
10 - 19	7	14.5	1.1614	8.1298
20 - 29	13	24.5	1.3892	18.0596
30 - 39	22	34.5	1.5378	33.8316
40 - 49	11	44.5	1.6484	18.1324
50 - 59	6	54.5	1.7364	10.4184
60 - 69	5	64.5	1.8096	9.0480
Total	64			97.6198

$\Sigma f . \log x = 97.6198$

$\Sigma f = 64$

Substituting these values in the formula. .

$$G.M. = \text{Antilog}\left[\frac{\Sigma f . \log x}{\Sigma f}\right]$$

$$= \text{Antilog}\left[\frac{97.6198}{64}\right]$$

$$= \text{Antilog}\ (1.5253)$$

$$G.M. = 33.52 \text{ units}$$

4.8 HARMONIC MEAN (H.M)

Harmonic mean is defined only for non-zero positive values. It is the reciprocal of mean of reciprocal of values. In symbols

$$H.M. = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$$

Example 4.8.1

If an investor buys shares of Rs.9000 at a price of Rs.45/- per share and shares of Rs.9000 at Rs.36/- per share. Calculate the average price per share.

Solution:

x = Price / shares

= 45, 36

$$H.M. = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$$

$$= \frac{2}{\frac{1}{45} + \frac{1}{36}}$$

= Rs. 40 / share.

4.10 HARMONIC MEAN OF A FREQUENCY DISTRIBUTION

It is a weighted harmonic mean in which weights are the frequencies.

$$\text{H. M.} = \frac{\Sigma f}{\Sigma(f/x)}$$

Example: 4.10.1

Calculate Harmonic Mean for the following frequency distribution.

C. I.	10 — 19	20 — 29	30 — 39	40 — 49	50 — 59	60 — 69
f	7	13	22	11	6	5

Solution:

C.I.	f	x	f / x
10 - 19	7	14.5	0.4828
20 - 29	13	24.5	0.5306
30 - 39	22	34.5	0.6377
40 - 49	11	44.5	0.2472
50 - 59	6	54.5	0.1101
60 - 69	5	64.5	0.0775
Total	64		2.0859

$$\text{H.M.} = \frac{\Sigma f}{\Sigma(f/x)}$$

$$= \frac{64}{2.0859} = 30.68$$

4.11 MEDIAN $(\tilde{x})$

Median is defined as the middle value of the data when the values are arranged in ascending or descending order.

If there are even number of values in the data, the mean of two middle values in the array is taken as median. Although mean is the most commonly used measure of central tendency, there are a number of situations in which the median is a better measure. The mean is influenced by extreme values in the data but the median is not. So whenever there are extreme values in the data the median is usually a better measure than mean. Likewise when the data are listed in order by ranks, the median is the appropriate measure.

Example: 4.11.1

Find the median for the following data.

3, 27, 6, 8, 2, 2, 3, 9, 6

Solution:

Arranging the values in ascending order, the array is

	Values	Position in the array
	2	1st
	2	2nd
	3	3rd
	3	4th
median	6	5th
	6	6th
	8	7th
	9	8th
	27	9th

Middle value = median = 6

Example : 4.11.2.

Find the median for the following data.

21, 17, 14, 62, 18, 26, 32, 20, 47, 30, 27, 32

Solution:

Arranging the values in ascending order, the array is

arranged values	Position in the array
14	1st
17	2nd
18	3rd
20	4th
21	5th
26	6th
27	7th
30	8th
32	9th
32	10th
47	11th
62	12th

Two middle values

$$\text{Median} = \frac{26 + 27}{2}$$

$$\tilde{x} = 26.5$$

4.12 MEDIAN OF A FREQUENCY DISTRIBUTION

In a frequency distribution actual values of the data are not known and hence arrangement of the data in an array to determine the middle value is not possible. The assumption that the values of the data are mid-points of the intervals cannot be made because median considers the positions of the values in the array instead of their numerical quantity. It is therefore more appropriate to assume that

> *Values of the data in an interval are evenly (or uniformly) spread in that interval*

This assumption helps in determining the numerical value of the observation at any specified position in the array.

Example: 4.12.1

There are 8 values in the interval 15 – 25.

Determine the 4th value in the array assuming that the values are evenly spread in the interval.

Solution:

The width of the interval is 10 and there are 8 values, equally spaced, in this interval, therefore the difference between the values is

$$\frac{\text{width of the interval}}{\text{No. of values in the interval}} = \frac{10}{8} = 1.25$$

Starting with the initial value of the interval, which is 15, the first value is assumed to be $15 + 1.25 = 16.25$

Accordingly the second value is $15 + 2(1.25) = 17.50$
the third value is $15 + 3(1.25) = 18.75$
and the fourth value is $15 + 4(1.25) = 20.00$

Hence the 4th value is 20.00

Example: 4.12.2

Find median for the frequency distribution given below:

C.B.	f
5 - 15	5
15 - 25	8
25 - 35	3
35 - 45	2
	18

Solution

C.B.	f	C.F. <	Positions of values
5 - 15	5	5	1st to 5th
15 - 25	8	13	6th to 13th
25 - 35	3	16	14th to 16th
35 - 45	2	18	17th & 18th

In a frequency distribution median is considered as the $\left(\frac{N}{2}\right)$th value in the array whether N is even or odd.

(Note: $N = \Sigma f$)

So Median = 9th value in the data array.

= 4th value in the interval (15 – 25)

= 15 + 4 (1.25) – (as in example 4.12.1

= 20.00

Based on the above reasoning, a formula to calculate the median of a frequency distribution has been derived as

$$\tilde{x} = l + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

l	=	lower boundary of median group
h	=	width of median group
f	=	frequency of median group
N	=	Σf
C	=	Cummulative frequency preceeding median group

An example to illustrate the application of this formula follows:

Example : 4.12.3

Calculate Median for the frequency distribution given on the

C. I.	f
30.0 — 34.9	4
35.0 — 39.9	11
40.0 — 44.9	17
45.0 — 49.9	16
50.0 — 54.9	14
55.0 — 59.9	12
60.0 — 64.9	8
65.0 — 69.9	5
70.0 — 74.9	5
75.0 — 79.9	3
80.0 — 84.9	2
85.0 — 89.9	2
90.0 — 94.9	1

Solution :

Class Boundaries	f	C.F <
29.95 — 34.95	4	4
34.95 — 39.95	11	15
39.95 — 44.95	17	32
44.95 — 49.95	16	48 ← c
l → 49.95 — 54.95	14 ← f	62 ← 50th value
54.95 — 59.95	12	74
59.95 — 64.95	8	82
64.95 — 69.95	5	87
69.95 — 74.95	5	92
74.95 — 79.95	3	95
79.95 — 84.95	2	97
84.95 — 89.95	2	99
89.95 — 94.95	1	100

$\left(\frac{N}{2}\right)$th $= \left(\frac{100}{2}\right)$th $=$ 50th value is the median.

50th value is in the interval 49.95 – 54.95.

This interval is called median group.

Now applying the formula

$$\tilde{x} = l + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

where median group is the interval (49.95 – 54.95)

l	=	49.95	the lower class boundary of median group
h	=	5	the width of median group
f	=	14	the class frequency of median group
N	=	100	the sum of frequencies
C	=	48	the cumulative frequncy of the interval preceeding median group.

$$\tilde{x} = 49.95 + \frac{5}{14}\left(\frac{100}{2} - 48\right)$$

$$= 49.95 + 0.714$$

$$= 50.664$$

4.13 PARTITION VALUES OR QUANTILES

Quartiles:

There are three values which can divide the arranged data in four equal parts or quarters.These values are called quartiles.

The first quartile, Q_1, is the value that is larger than one quarter of the values.

The second quartile, Q_2, is the value that is larger than one half of the values.

The third quartile, Q_3, is the value that is larger than three quarters of the values.

Example : 4.13.1

For the data of example 2.2.1 find the three quartiles.

Solution :

The data is arranged in ascending order as :

5,5,5,5,5,5,6,6,6,7 ↓ Q_1 (10 Values) | 7,7,8,8,8,8,8,8,8,8, ↓ Q_2 (10 Values) | 9,9,9,10,10,10,10,10,10,11 ↓ Q_3 (10 Values) | 11,12,12,12,13,13,13,14,14,15 (10 Values)

$$Q_1 = \frac{7+7}{2} = 7$$

$$Q_2 = \frac{8+9}{2} = 8.5$$

$$Q_3 = \frac{11+11}{2} = 11$$

Deciles & percentiles

Nine values can divide the arranged data in ten equal parts. These values are called deciles and denoted by $D_1, D_2, \ldots, D_9$. Similarly the 99 values which divide the arranged data in 100 equal parts are called percentiles and denoted by

$$P_1, P_2, \ldots P_{99}.$$

Clearly $Q_2 = D_5 = P_{50} =$ Median.

For a frequency distribution these quantities called partition values or quantiles are calculated using a formula which is a generalization of the formula for the median.

$$Q_i = l + \frac{h}{f}\left(\frac{iN}{4} - C\right) ; \quad i = 1,2,3$$

$$D_i = l + \frac{h}{f}\left(\frac{iN}{10} - C\right) ; \quad i = 1,2,3 \ldots, 9$$

$$P_i = l + \frac{h}{f}\left(\frac{iN}{100} - C\right) ; \quad i = 1,2,3,\ldots ,99$$

Note :

$Q_1 = P_{25}$ $\quad D_1 = P_{10}$

$Q_2 = P_{50}$ $\quad D_2 = P_{20}$

$Q_3 = P_{75}$ $\quad D_9 = P_{90}$ and so on

Example . 4.13.2 :

For the following frequency distribution calculate Q_1, D_4, and P_{95}

C.I.	10 — 19	20 — 29	30 — 39	40 — 49	50 — 59	60 — 69
f	7	13	22	11	6	5

$$Q_2 = \frac{8 + 9}{2} = 8.5$$

$$Q_3 = \frac{11 + 11}{2} = 11$$

Deciles & percentiles

Nine values can divide the arranged data in ten equal parts. These values are called deciles and denoted by $D_1, D_2, ---, D_9$. Similarly the 99 values which divide the arranged data in 100 equal parts are called percentiles and denoted by

$$P_1,\ P_2,\ ---\ P_{99}.$$

Clearly $Q_2 = D_5 = P_{50} =$ Median.

For a frequency distribution these quantities called partition values or quantiles are calculated using a formula which is a generalization of the formula for the median.

$$Q_i = l + \frac{h}{f}\left(\frac{iN}{4} - C\right) \; ; \quad i = 1,2,3$$

$$D_i = l + \frac{h}{f}\left(\frac{iN}{10} - C\right) \; ; \quad i = 1,2,3 \ldots, 9$$

$$P_i = l + \frac{h}{f}\left(\frac{iN}{100} - C\right) \; ; \quad i = 1,2,3,\ldots,99$$

Note :

$Q_1 = P_{25}$ $\quad D_1 = P_{10}$

$Q_2 = P_{50}$ $\quad D_2 = P_{20}$

$Q_3 = P_{75}$ $\quad D_9 = P_{90}$ and so on

Example . 4.13.2 :

For the following frequency distribution calculate Q_1, D_4, and P_{95}

C.I.	10—19	20—29	30—39	40—49	50—59	60—69
f	7	13	22	11	6	5

$$Q_2 = \frac{8 + 9}{2} = 8.5$$

$$Q_3 = \frac{11 + 11}{2} = 11$$

Deciles & percentiles

Nine values can divide the arranged data in ten equal parts. These values are called deciles and denoted by $D_1, D_2, ---, D_9$. Similarly the 99 values which divide the arranged data in 100 equal parts are called percentiles and denoted by

$$P_1,\ P_2,\ ---\ P_{99}.$$

Clearly $Q_2 = D_5 = P_{50} =$ Median.

For a frequency distribution these quantities called partition values or quantiles are calculated using a formula which is a generalization of the formula for the median.

$$Q_i = l + \frac{h}{f}\left(\frac{iN}{4} - C\right) \ ; \quad i = 1, 2, 3$$

$$D_i = l + \frac{h}{f}\left(\frac{iN}{10} - C\right) \ ; \quad i = 1, 2, 3 \ldots\ldots, 9$$

$$P_i = l + \frac{h}{f}\left(\frac{iN}{100} - C\right) \ ; \quad i = 1, 2, 3, \ldots\ldots, 99$$

Note :

$Q_1 = P_{25}$ $\quad D_1 = P_{10}$

$Q_2 = P_{50}$ $\quad D_2 = P_{20}$

$Q_3 = P_{75}$ $\quad D_9 = P_{90}$ and so on

Example . 4.13.2 :

For the following frequency distribution calculate Q_1, D_4, and P_{95}

C.I.	10 — 19	20 — 29	30 — 39	40 — 49	50 — 59	60 — 69
f	7	13	22	11	6	5

Solution :

C. I.	C. B.	f	C.F. <	
10 – 19	9.5 — 19.5	7	7	
20 – 29	19.5 — 29.5	13	20	← 16th value
30 – 39	29.5 — 39.5	22	42	← 25.6th value
40 – 49	39.5 — 49.5	11	53	
50 – 59	49.5 — 59.5	6	59	
60 – 69	59.5 — 69.5	5	64	← 60.8th value
Total		64		

Q_1 is $\left(\frac{N}{4}\right)$ th or 16th value.

16th value is in the interval (19.5 – 29.5)

$\therefore$ Group of Q_1 is (19.5 – 29.5)

$$Q_1 = l + \frac{h}{f}\left(\frac{N}{4} - C\right)$$

$$= 19.5 + \frac{10}{13}(16 - 7)$$

$$= 19.5 + 6.9$$

~~$Q_1 = 26.4$~~

D_4 is $\left(\frac{4N}{10}\right)$ th value or (25.6 th) value (between 25th & 26th values)

Group of D_4 is (29.5 – 39.5)

$$\therefore \quad D_4 = l + \frac{h}{f}\left(\frac{4N}{10} - C\right)$$

$$= 29.5 + \frac{10}{22}(25.6 - 20)$$

$$= 29.5 + 2.5$$

$$D_4 = 32.0$$

P_{95} is $\left(\frac{95N}{100}\right)$ th value or (60.8) th value (between & 61st value) Group of P_{95} is (59.6 – 69.5)

$$\therefore \quad P_{95} = l + \frac{h}{f}\left(\frac{95N}{100} - C\right)$$

$$= 59.5 + \frac{10}{5}(60.8 - 59)$$

$$= 59.5 + 3.6$$

$$P_{95} = 63.1$$

4.16 MODE OF A FREQUENCY DISTRIBUTION

When raw data are converted into a frequency distribution, a class interval with maximum frequency is defined as a modal class. A frequency distribution can have a modal class irrespective of the existence of mode in ungrouped data. So whenever mode is to be used as a measure of central tendency, it should be calculated from grouped data.

> *In a frequency distribution mode is that value of the variable for which the frequency curve takes maximum height.*

A frequency distribution with one mode is called unimodal and with two modes it is called a bimodal frequency distribution.

The formula used to find mode in a frequency distribution is

$$\hat{x} = l + \left[\frac{f_m - f_1}{2f_m - f_1 - f_2}\right] \times h$$

l = Lower boundary of modal group.
h = Width of modal group
f_m = Maximum frequency
f_1 = Frequency preceeding modal group.
f_2 = Frequency following modal group.

Example: 4.16.1

Calculate mode for the frequency distribution of example 4.13.2

Solution :

Class interval with maximum frequency is the modal class. 22 is the maximum frequency therefore (29.5 – 39.5) is the modal class.

C. I.	f	C.B
10 - 19	7	9.5 — 19.5
20 - 29	13	19.5 — 29.5
30 - 39	22	29.5 — 39.5
40 - 49	11	39.5 — 49.5
50 - 59	6	49.5 — 59.5
60 - 69	5	59.5 — 69.5

$$\hat{x} = l + \left[\frac{f_m - f_1}{2f_m - f_1 - f_2}\right] \times h$$

$$= 29.5 + \left[\frac{22 - 13}{44 - 13 - 11}\right] \times 10$$

$$= 29.5 + 4.5$$

$$\hat{x} = 34.0$$

CHAPTER 2

MEASURES OF DISPERSION

5.1 INTRODUCTION

Dispersion is another important characteristic of data which describes the extent to which the observations vary among themselves.

The following data of three different samples have the same means but their dispersions are different.

Data 1 : 3, 27, 6, 8, 2, 2, 3, 9, 6, 4 ; $\bar{x}_1 = 7$

Data 2 : 7, 8, 9, 7, 7, 6, 5, 9, 5, 7, ; $\bar{x}_2 = 7$

Data 3 : 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ; $\bar{x}_3 = 7$

The dispersion of data 3 is clearly zero.

The dispersion of data 1 is greater than the dispersion of data 2.

The dispersion is defined as the scatter or spread of the values from one another or from some common value.

The methods to compute the amount of dispersion present in any data are called "Measures of Dispersion" or "Measures of variation"

There are four measures of dispersion :

1. Range
2. Quartile Deviation.
3. Mean Deviation.
4. Standard Deviation.

5.2 RANGE (R)

Range is the simplest measure of dispersion and is defined as the difference between the maximum and minimum values of the data, that is

$$R = X_{max} - X_{min}$$

Range is a rough and crude measure as it ignores the variation among all the values.

In a frequency distribution range is the difference between upper class boundary of the last interval and lower class boundary of the first interval.

5.3 QUARTILE DEVIATION (QD)

The difference between third and first quartiles is called Interquartile Range.

Quartile Deviation is half of Interquartile Range also known as Semi Interquartile Range.

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

Example 5.3.1 :

Calculate Quartile Deviation for the following data.

16, 13, 12, 18, 9, 9, 16, 3, 7, 13

2, 13, 8, 20, 17, 16, 9, 7, 14, 18

Solution:

The arrangement of data in ascending order is

2, 3, 7, 7, 8	9, 9, 9, 12, 13	13, 13, 14, 16, 16	16, 17, 18, 18, 20

$$Q_1 = \frac{8+9}{2} = 8.5 \qquad Q_2 = \frac{13+13}{2} = 13 \qquad Q_3 = \frac{16+16}{2} = 16$$

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

$$= \frac{16 - 8.5}{2} \quad \text{or} \quad Q.D. = 3.75$$

Example 5.3.2 :

For the frequency distribution given below, calculate Q.D.

C.I	f
2.5 — 2.9	2
3.0 — 3.4	7
3.5 — 3.9	17
4.0 — 4.4	25
4.5 — 4.9	20
5.0 — 5.4	12
5.5 — 5.9	9
6.0 — 6.4	8

Solution:

	C.B	f	C.F<	
	2.45 — 2.95	2	2	
	2.95 — 3.45	7	9	
Group of Q_1 →	3.45 — 3.95	17	26	← 25th value
	3.95 — 4.45	25	51	
	4.45 — 4.95	20	71	
Group of Q_3 →	4.95 — 5.45	12	83	← 75th value
	5.45 — 5.95	9	92	
	5.95 — 6.45	8	100	

Q_1 is $\left(\frac{100}{4}\right)$ *th* or 25th value; Q_3 is $\left(\frac{3 \times 100}{4}\right)$ *th* or 75th

Group of Q_1 is **3.45 – 3.95**

Now

$$Q_1 = l + \frac{h}{f}\left(\frac{N}{4} - c\right)$$

$$= 3.45 + \frac{0.5}{17}(25 - 9)$$

$$= 3.45 + 0.47$$

$$= 3.92 \text{ units}$$

Group of Q_3 is **4.95 – 5.45**

Now

$$Q_3 = l + \frac{h}{f}\left(\frac{3N}{4} - c\right)$$

$$= 4.95 + \frac{0.5}{12}(75 - 71)$$

$$= 4.95 + 0.17$$

$$= 5.12 \text{ units}$$

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{5.12 - 3.92}{2} = 0.6 \text{ units}$$

MEAN DEVIATION (M.D)

Dispersion can be measured in terms of the quantities that each value of the data deviates from average value. If deviations of the values are taken from mean, the mean of these deviations vanishes as "the sum of deviations from mean is always zero". (See example 4.3.1).

Absolute deviations are therefore used to find the mean of deviations called Mean Absolute Deviation or simply Mean Deviation. Thus Mean Deviation is the sum of absolute deviations from mean divided by the number of values.

$$M.D. = \frac{\Sigma|x - \bar{x}|}{n}$$

Median is sometimes used to find the deviations of the values. The mean of these absolute deviations is called mean deviation from median. Thus

$$M.D.(\tilde{x}) = \frac{\Sigma|x - \tilde{x}|}{n}$$

Example 5.4.1:

For the data given on the next page, calculate

(i) Mean Deviation

(ii) Mean Deviation from median

16	13	2	18	9
9	16	3	7	13
12	13	8	20	17
16	9	7	14	18

Solution:

$$\bar{x} = \frac{240}{20}$$

$$\boxed{\bar{x} = 1?\ \text{units}}$$

For median arrange the values in an array and pick the two middle values.

Array: 2, 3 , 7, 7, 8, 9, 9, 9, 12, [13, 13,] 13, 14, 16, 16, 16, 17, 18, 18, 20

middle values

$$\tilde{x} = \frac{13 + 13}{2}$$

$$\boxed{\tilde{x} = 13\ \text{units}}$$

The deviations from mean ($\bar{x} = 12$) and from median ($\tilde{x} = 13$) are shown on the next page in a tabular form.

$$M.D. = \frac{\Sigma|x - \bar{x}|}{n} \qquad M.D.\ (\tilde{x}) = \frac{\Sigma|x - \tilde{x}|}{n}$$

$$= \frac{84}{20} \qquad = \frac{81}{20}$$

$$= 4.2\ \text{units} \qquad = 4.05\ \text{units}$$

$$\bar{x} = 4.43$$

$$M.D.(\bar{x}) = \frac{\Sigma f.|x - \bar{x}|}{\Sigma f} = \frac{70.46}{100} = 0.7046$$

5.6 STANDARD DEVIATION (s or σ)

Standard Deviation is the most widely used measure of dispersion. It is defined as the positive square root of a quantity called variance.

Variance of a sample of n values, called the sample variance and denoted by s^2, is obtained by the formula

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1} \quad ; \text{ Sample Variance}$$

Variance of a population of N values, called the population variance and denoted by σ^2, is obtained by the formula

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N} \quad ; \text{ Population Variance}$$

The sample and population standard deviations are obtained by taking positive square roots of respective variances:

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}} \quad ; \text{ Sample Standard Deviation}$$

$$\sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}} \quad ; \text{ Population Standard Deviation.}$$

Example 5.6.1

Calculate variance and standard deviation for the following sample data:

23, 17, 14, 62, 18, 26, 32, 20,47, 30, 27, 32

Solution :

The computations are presented in a tabular form in table 5.6.1

Table 5.6.1

x	$(x-\bar{x})$ $\bar{x} = 29$	$(x-\bar{x})^2$
23	−6	36
17	−12	144
14	−15	225
62	+33	1089
18	−11	121
26	−3	9
32	+3	9
20	−9	81
47	+18	324
30	+1	1
27	−2	4
32	+3	9
348	0	2052

$$\bar{x} = \frac{\Sigma x}{n} = \frac{348}{12} = 29 \text{ units}$$

Variance

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{2052}{12-1} = 186.55 \text{ Squared units}$$

Standard Deviation :-

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\text{Variance}} = \sqrt{186.55} = 13.66 \text{ units}$$

To minimize the round off errors in the calculation of variance when the mean is not a whole number, another formula is used, called computational form of variance.

$$s^2 = \frac{n\Sigma x^2 - (\Sigma x)^2}{n(n-1)}$$ Computational form of sample variance

$$\sigma^2 = \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2$$; Computational form of population variance

The standard deviations in computational forms are

$$s = \sqrt{\frac{n\Sigma x^2 - (\Sigma x)^2}{n(n-1)}}$$

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

Example 5.6.2

For the sample data of example 5.6.1, calculate variance and standard deviation using computational formula.

x	x^2
23	529
17	289
14	196
62	3844
18	324
26	676
32	1024
20	400
47	2209
30	900
27	729
32	1024
348	12144

Variance

$$s^2 = \frac{n\Sigma x^2 - (\Sigma x)^2}{n(n-1)}$$

$$= \frac{12(12144) - (348)^2}{12(12-1)}$$

$$= \frac{145728 - 121104}{12 \times 11}$$

$$= \frac{24624}{132}$$

$= 186.55$ squared units

Standard Deviation

$$s = \sqrt{Variance}$$

$$= \sqrt{186.55}$$

$= 13.66$ units

5.7 STANDARD DEVIATION FOR A FREQUENCY DISTRIBUTION

In a frequency distribution the variance is calculated by the formula

$$s^2 = \frac{\Sigma f(x-\bar{x})^2}{\Sigma f} \quad \text{Definition form}$$

$$\text{or } s^2 = \frac{\Sigma f x^2}{\Sigma f} - \left(\frac{\Sigma f x}{\Sigma f}\right)^2 \quad \text{Computational form}$$

and hence the standard deviation is

$s = \sqrt{\dfrac{\Sigma f(x-\bar{x})^2}{\Sigma f}}$	Definition form
or $s = \sqrt{\dfrac{\Sigma fx^2}{\Sigma f} - \left(\dfrac{\Sigma fx}{\Sigma f}\right)^2}$	Computational form

Computational form of variance and standard deviation are easier to apply when the mean of the distribution is not an exact number.

If frequency distribution represents a population, the variance and standard deviation are obtained by the above formula but s^2 is replaced by σ^2 and s by σ

Example 5.7.1

For the frequency distribution of example 5.3.2., calculate mean, variance and standard deviation.

Solution :

Table 5.7.1 Computation for Variance

C.I	f	x	f.x	$(x-\bar{x})$	$(x-\bar{x})^2$	$f.(x-\bar{x})^2$	fx^2
2.5 - 2.9	2	2.7	5.4	− 1.83	3.3489	6.6978	14.58
3.0 - 3.4	7	3.2	22.4	− 1.33	1.7689	12.3823	71.68
3.5 - 3.9	17	3.7	62.9	− 0.83	0.6889	11.7113	232.73
4.0 - 4.4	25	4.2	105.0	− 0.33	0.1089	2.7225	441.00
4.5 - 4.9	20	4.7	94.0	+0.17	0.0289	0.5780	441.80
5.0 - 5.4	12	5.2	62.4	+0.67	0.4489	5.3868	324.48
5.5 - 5.9	9	5.7	51.3	+1.17	1.3689	12.3201	292.41
6.0 - 6.4	8	6.2	49.6	+1.67	2.7889	22.3112	307.52
	100	-	453.0		-	74.1100	2126.20

$$\bar{x} = \frac{453.0}{100} = 4.53$$

Definition form

$$s^2 = \frac{\Sigma f(x-\bar{x})^2}{\Sigma f}$$

Computational form

$$s^2 = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2$$

$$= \frac{74.11}{100}$$

$$= 0.7411$$

$$s = \sqrt{0.7411}$$

$$= 0.86 \text{ units}$$

$$= \frac{2126.2}{100} - \left(\frac{453}{100}\right)$$

$$= 21.262 - 20.5209$$

$$s = \sqrt{0.7411}$$

$$= 0.86 \text{ units}$$

Note : Computational form for variance is obviously simpler to use and hence is preferred over the definition form.

Example 5.8.1

For the frequency distribution of example 5.3.2 calculate variance using method of codes.

Solution :

Table 5.8.1 :

codes

C.I	f	x	u	f.u	fu²
2.5 – 2.9	2	2.7	−3	−6	18
3.0 – 3.4	7	3.2	−2	−14	28
3.5 – 3.9	17	3.7	−1	−17	17
4.0 – 4.4	25	4.2	0	0	0
4.5 – 4.9	20	4.7	1	20	20
5.0 – 5.4	12	5.2	2	24	48
5.5 – 5.9	9	5.7	3	27	81
6.0 – 6.4	8	6.2	4	32	128
Total	100			66	340

$$S_u^2 = \frac{\Sigma f u^2}{\Sigma f} - \left(\frac{\Sigma f u}{\Sigma f}\right)^2$$

$$= \frac{340}{100} - \left(\frac{66}{100}\right)^2$$

$$= 3.4 - 0.4356$$

$$= 2.9644$$

$$s_x^2 = (0.5)^2 . (2.9644)$$

$$= 0.7411 \text{ squared units}$$

CHAPTER 3

7.3 MEASURES OF SKEWNESS

The "*Skewness*" is used as a term opposite to "*Symmetry*".

To measure skewness is to measure the extent to which and also the direction in which the distribution (or curve) is non-symmetrical or skewed.

The figures 7.3 below show curves of symmetrical and skewed distributions.

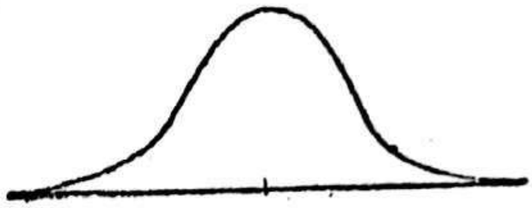
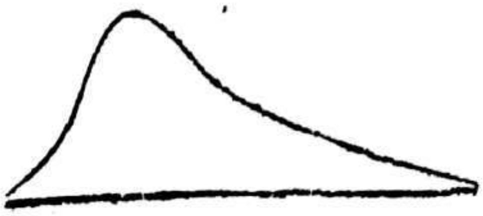


Fig 7.3

(a) Symmetrical



(b) Right skewed (c) Left skewed

There are three measures of skewness defined as

1. The difference between mean and mode, that is

   Skewness = mean − mode

or $Sk = \bar{x} - \hat{x}$

In a moderately skewed distribution, the empirical relation between $\bar{x}$, $\tilde{x}$ and $\hat{x}$ is

$$(\bar{x} - \hat{x}) = 3(\bar{x} - \tilde{x})$$

Therefore $Sk = 3(\bar{x} - \tilde{x})$

2. The difference between the distances (or differences)

of Q_3 and Q_2 and Q_2 and Q_1, that is

$$Sk = (Q_3 - Q_2) - (Q_2 - Q_1)$$

or $$Sk = Q_3 - 2Q_2 + Q_1$$

3. The third order moment about mean, that is

$$Sk = m_3 \quad or \quad Sk = \mu_3$$

The three relative measures of skewness, called coefficient of skewness are :

1. $$C.Sk = \frac{3(\bar{x} - \tilde{x})}{s} = \frac{\bar{x} - \hat{x}}{s}$$

2. $$C.Sk = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

3. $$C.Sk = \beta_1 = \frac{m_3^2}{m_2^3} \quad or \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

(with sign of m_3 or μ_3)

If a distribution is symmetrical, the measure of skewness must be zero.

Thus

Skewness = 0 implies symmetry.

Example 7.3.1

For the frequency distribution of example 5.3.2, calculate the three coefficients of skewness.

Solution:

(1) Mean, median and standard deviation of the distribution have already been calculated in examples 5.5.1 and 5.7.1 as :

$$\bar{x} = 4.53 \;,\; \tilde{x} = 4.43 \quad \text{and} \quad s = 0.86$$

Therefore the coefficient of skewness is

$$C.Sk = \frac{3(4.53 - 4.43)}{0.86}$$

$$= 0.3488 \;,\; \cong 0.35$$

(2) Q_1 and Q_3 of the distribution are calculated in example 5.3.2 as

$$\boxed{Q_1 = 3.92 \text{ and } Q_3 = 5.12}$$

Q_2 or median is given in example 5.5.1 as $Q_2 = 4.43$

The coefficient of skewness is thus

$$C.Sk = \frac{5.12 + 3.92 - 2(4.43)}{5.12 + 3.92}$$

$$= \frac{0.18}{9.04} = 0.0199 \cong 0.02$$

3) Results of example 7.2.1 give

$\boxed{m_2 = 0.74}$ and $\boxed{m_3 = 0.16}$

The coefficient of skewness is thus

$$\beta_1 = \frac{(0.16)^2}{(0.74)^3} = \frac{0.0256}{0.405224} = 0.063175 \text{ or } 0.06$$

8.7 SOME SPECIAL WEIGHTED AGGREGATIVE INDICES

Measuring the "relative importance" or "weights" of the commodities is a difficult task. To avoid this difficulty of measurement, physical quantities of the commodities either consumed or produced during the base or current time periods are taken as the measures of weights.

1. Laspeyre's price index :

If the quantities of base time period are taken as weights, the weighted aggregative price index is called Laspeyre's price index.

$$I_L = \frac{\Sigma P_n . Q_o}{\Sigma P_o . Q_o} \times 100$$

2. Paasche's price index:

If the quantities of current time period are taken as weights, the weighted aggregative index is called Paasche's price index.

$$I_P = \frac{\Sigma P_n . Q_n}{\Sigma P_0 . Q_n} \times 100$$

3. Fisher's price index

The geometric mean of Laspeyre's and Paasche's price indices is called Fisher's price index.

Fisher's price index is also a weighted aggregative price index because it is an average (G.M.) of two weighted aggregative indices.

$$I_F = \sqrt{I_L . I_P}$$

or

$$I_F = \sqrt{\frac{\Sigma P_n . Q_o}{\Sigma P_o . Q_o} \cdot \frac{\Sigma P_n . Q_n}{\Sigma P_o . Q_n}} \times 100$$

Marshall's Price Index

$$I_M = \frac{\sum PnQo + \sum PnQn}{\sum PoQo + \sum PoQn} \times 100$$

Example 8.7.1

Calculate Laspeyre's Paasche's and Fisher's index numbers for the given data : Base = 1975.

Commodity	Price		Quantity	
	1975	1976	1975	1976
	P_o	P_1	Q_o	Q_1
Petrol	6.60	7.10	240	330
Diesel	4.15	4.90	185	210
Gas	1.25	2.00	315	345
Kerosene Oil	0.65	1.30	260	115

Solution :

a) Laspeyre's price index : 1975 quantities as weights.

Commodity	Weighted Price	
	$P_o Q_o$	$P_1 Q_o$
Petrol	1584.00	1704.00
Diesel	767.75	906.50
Gas	393.75	630.00
Kerosene	169.00	338.00
	2914.50	3578.50

$$I_L = \frac{3578.50}{2914.50} \times 100 = 122.78\ \%$$

b) Paasche's price index : 1976 quantities as weights.

Commodity	Weighted Price	
	$P_o Q_1$	$P_1 Q_1$
Petrol	2178.00	2343.00
Diesel	871.50	1029.00
Gas	431.25	690.00
Kerosene	74.75	149.50
	3555.50	4211.50

$$I_P = \frac{4211.50}{3555.50} \times 100 = 118.45\ \%$$

c) Fisher's price index

$$I_F = \sqrt{I_L \times I_P}$$

$$= \sqrt{122.78 \times 118.45} = 120.60\ \%$$

The results are tabulated below:

Year	Index Number		
	I_L	I_P	I_F
1975	100.00	100.00	100.00
1976	122.78	118.45	120.60

Example 8.7.2

Construct Laspeyre's, Paasche's and Fisher's price indices using 1954 as base year.

Commodity	Price				Quantity			
	1954	1964	1974	1984	1954	1964	1974	1984
	P_o	P_1	P_2	P_3	Q_o	Q_1	Q_2	Q_3
Bread	0.12	0.25	1.50	3.25	25	32	40	52
Meat	2.50	4.00	9.00	21.00	20	28	35	46
Fish	3.00	5.50	12.50	28.7[illegible]	3	4	4	7
Butter	5.00	7.25	11.00	18.50	8	10	15	22
Milk	1.50	2.25	4.50	6.00	12	17	24	34
Tea	2.75	5.75	13.25	20.50	5	8	11	20

Solution :

For Laspeyre's indices :

Commodity	Weighted Price			
	$P_o.Q_o$	$P_1.Q_o$	$P_2.Q_o$	$P_3.Q_o$
Bread	3.00	6.25	37.50	81.25
Meat	50.00	80.00	180.00	420.00
Fish	9.00	16.50	37.50	86.25
Butter	40.00	58.00	88.00	148.00
Milk	18.00	27.00	54.00	72.00
Tea	13.75	28.75	66.25	102.50
Total	133.75	216.50	463.25	910.00

$$I_{1964} = \frac{216.50}{133.75} \times 100 = 162\ \%$$

$$I_{1974} = \frac{463.25}{133.75} \times 100 = 346\ \%$$

$$I_{1984} = \frac{910}{133.75} \times 100 = 680\ \%$$

For Paasche's indices :

Commodity	Weighted Price					
	$P_o Q_1$	$P_1 Q_1$	$P_o Q_2$	$P_2 Q_2$	$P_o Q_3$	$P_3 Q_3$
Bread	3.84	8.00	4.80	60.00	6.24	169.00
Meat	70.00	112.00	87.50	315.00	115.00	966.00
Fish	12.00	22.00	12.00	50.00	21.00	201.25
Butter	50.00	72.50	75.00	165.00	110.00	407.00
Milk	25.50	38.25	36.00	108.00	51.00	204.00
Tea	22.00	46.00	30.25	145.75	55.00	410.00
Total	183.34	298.75	245.55	843.75	358.24	2357.25

$$I_{1964} = \frac{298.75}{188.34} \times 100 = 158.6\ \%$$

$$I_{1974} = \frac{843.75}{245.55} \times 100 = 343.6\ \%$$

$$I_{1984} = \frac{2357.25}{358.24} \times 100 = 658.0\ \%$$

For Fisher's indices

$$I_{1964} = \sqrt{162 \times 158.6} = 160.3\ \%$$

$$I_{1974} = \sqrt{346 \times 343.6} = 344.8\ \%$$

$$I_{1984} = \sqrt{680 \times 658} = 668.9\ \%$$

Results

Years	Index		
	Laspayre	Paasche	Fisher
1954	100	100.0	100.0
1964	162	158.6	160.3
1974	346	343.6	344.8
1984	680	658.0	668.9

CHAPTER 5

SIMPLE LINEAR REGRESSION

18.1 INTRODUCTION

This chapter introduces a technique which determines a relationship between two variables to estimate one of the variables for a given value of the other variable. The variable whose value is to be estimated is called dependent variable (y) whereas the variable whose value is given is called independent variable (x).

The following are few examples of pairs of dependent and independent variables:

	Independent variable (x)	Dependent variable (y)
1.	Price	Demand
2.	Rainfall	Yield
3.	Earning per share	Number of shares sold
4.	Research expenditures	Returns to the firm
5.	Credit sales	Bad debts
6.	Volume of production	Manufacturing expenses
7.	Expenses on protective measures	Compensations paid
8.	Workers age	Absenteeism

The type of relationship between the two variables that is the concern of this chapter is the linear or straight line relationship.

18.2 LINEAR REGRESSION MODEL

For a fixed value of the independent variable x, if the value of dependent variable y is observed a large number of times, different

18.7 TWO REGRESSION EQUATIONS FOR INTERDEPENDENT VARIABLES

Sometimes the two variables depend on each other called interdependent variables. Two regression lines can then be determined to estimate each of the variables by assuming the other variable independent.

Let x and y be two interdependent variables.

Assuming x to be independent, y can be estimated by the regression equation of y on x given by

$$\hat{y} = a + bx$$

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

and $$a = \bar{y} - b.\bar{x}$$

Again assuming y to be independent, x can be estimated by an analogous regression equation of x on y give by

$$\hat{x} = c + dy$$

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

and $$c = \bar{x} - d\bar{y}$$

Note that:

1. The signs of both the regression coefficients (b and d) must always be same.
2. Both the regression lines must pass through the point $(\bar{x}, \bar{y})$ which is the point of intersection of the two lines.
3. Coefficient of determination for both the lines is same and can very easily be found by multiplying the two regression coefficients, that is

$$r^2 = b \times d$$

4. If $r^2 = 1$ the two regression lines coincide and when $r^2 = 0$, the two regression lines are perpendicular.

Example 18.7.1

The heights and weights of five men are as follows:

Height (inches) : X	64	68	70	72	74
Weight (pounds) : Y	160	170	180	190	195

a) Determine the two lines of regression.

b) Estimate y for $x = 69$ inches and x for $y = 185$ pounds.

c) Determine the coefficient of determination.

Solution:

x	y	xy	x^2	y^2
64	160	10240	4096	25600
68	170	11560	4624	28900
70	180	12600	4900	32400
72	190	13680	5184	36100
74	195	14430	5476	38025
348	895	62510	24280	161025

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

$$= \frac{5 \times 62510 - 348 \times 895}{5 \times 24280 - (348)^2}$$

$$b = 3.68$$

$$a = \bar{y} - b\bar{x}$$

$$= \frac{895}{5} - (3.68)\left(\frac{348}{5}\right)$$

$$a = -77.13$$

$$\hat{y} = a + bx$$

$$\hat{y} = -77.13 + 3.68x$$

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

$$= \frac{5 \times 62510 - 348 \times 895}{5 \times 161025 - (895)^2}$$

$$d = 0.2659$$

$$c = \bar{x} - d\bar{y}$$

$$= \frac{348}{5} - 0.2659\left(\frac{895}{5}\right)$$

$$c = 22.0039$$

$$\hat{x} = c + dy$$

$$\hat{x} = 22.0039 + 0.2659y$$

for $x = 69$ inches

$\hat{y} = -77.13 + 3.68 \times 69$

$\hat{y} = 176.79$ pounds

for $y = 185$ pounds

$\hat{x} = 22.0039 + 0.2659 \times 185$

$\hat{x} = 71.2$ inches

Coefficient of determination r^2 is

$$r^2 = b \times d$$
$$= 3.68 \times 0.2659$$
$$= 0.9785$$

or $r^2 = 97.85\ \%$

18.8 INFERENCES ABOUT REGRESSION EQUATION

For two variables x and y, an observation of the variable y for a fixed x is

$$y = \mu_{y/x} + \in$$

Where $\mu_{y/x}$, the conditional mean of y given x, is expressed by the linear regression equation

$$\mu_{y/x} = \alpha + \beta x$$

If x contributes no information for the estimation (or prediction) of y, the $\mu_{y/x}$ does not change as x changes and regardless of the value of x same y-value will be estimated or predicted. This means that the regression coefficient β is zero. Therefore, to test the null hypothesis that there is no linear regression between x and y, the following null and alternative hypotheses are tested.

$$H_o : \beta = 0$$

$$H_A : \beta \neq 0 \quad or \quad \beta > 0 \quad or \quad \beta < 0$$

The sampling distribution of sample regression coefficient b is needed to test such hypothesis or to find confidence interval for β. But some assumptions which must be satisfied regarding the

CHAPTER 6

STATISTICAL INFERENCE : CHI-SQUARE TESTS

17.1 THE CHI-SQUARE DISTRIBUTION

Let x is a normally distributed random variable with mean μ and standard deviation σ. Then the standard normal variable z is given by

$$z = \frac{x - \mu}{\sigma}$$

Now, square and obtain

$$z^2 = \frac{(x - \mu)^2}{\sigma^2}$$

The probability distribution of the random variable z^2 is known as chi-square distribution with one degree of freedom written as $\chi^2_{(1)}$. Now take two independent, normally distributed random variables x_1 and x_2, standardize each, then square and sum;

$$z_1^2 + z_2^2 = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2$$

The distribution of $(z_1^2 + z_2^2)$ is chi-square with two degrees of freedom $\chi^2_{...}$

In general for n independent normal random variables, "the sum of squares of standard normal random variables" is

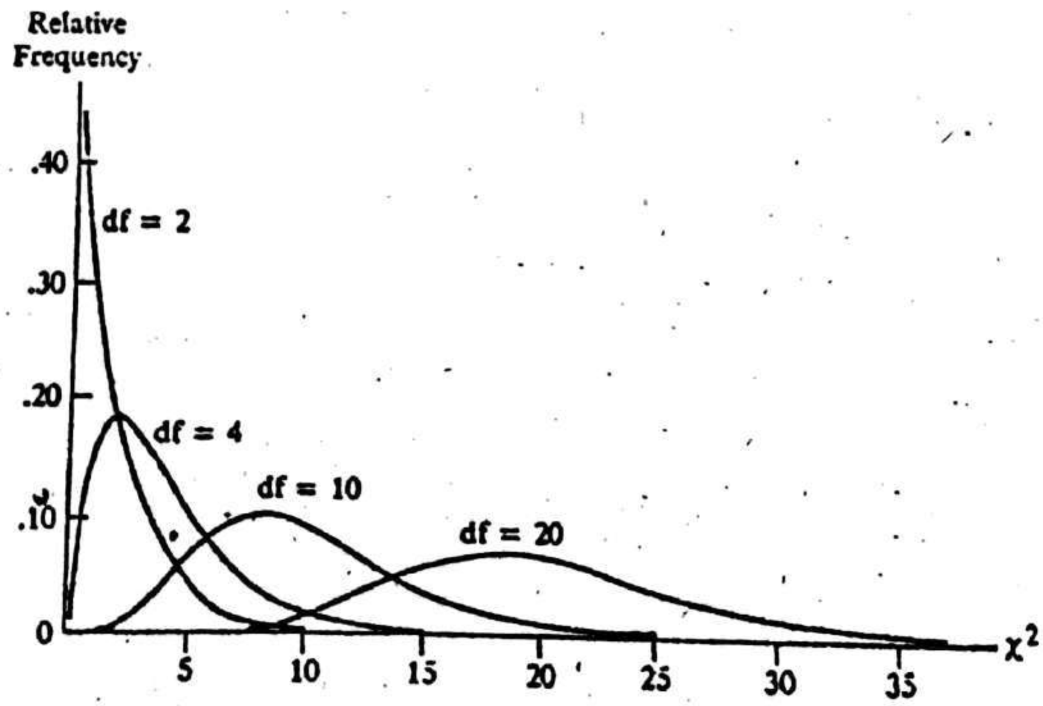
$$\sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$$

The distribution of Σz_i^2 is chi-square with n degrees of freedom, $\chi^2_{(n)}$

17.2 PROPERTIES OF CHI-SQUARE DISTRIBUTIONS

1. The value of χ_2 lies between 0 and ∞
2. The graph of a chi-square distribution is skewed and depends on the number of "degrees of freedom", which is the only parameter of the distribution and is denoted by ν

Four chi-square distributions with 2, 4, 10 and 20 degrees of freedom are sketched below:



3. The mean and variance of a chi-square distribution are given as

$$E(\chi^2) = \nu$$
$$V(\chi^2) = 2\nu$$

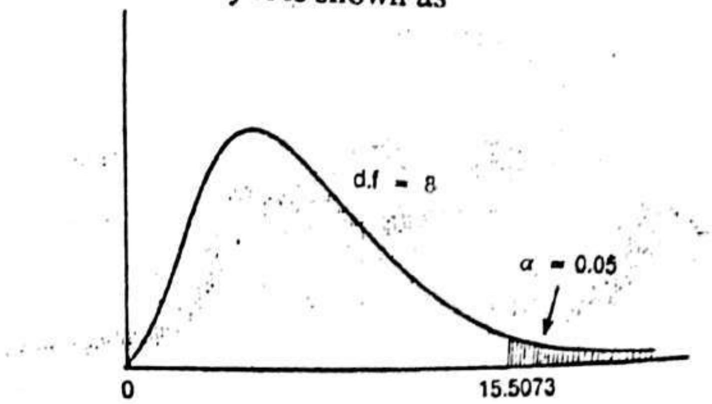
4. As number of degrees of freedom increases, the chi-square curves get more bell shaped and approach the normal curve but remember that a chi-square curve starts at zero, not at $-\infty$.

5. The values of χ^2 are given in table D

The columns of this table identify the value of α, area to the right of a chi-square value, and the rows correspond to the degrees of freedom ν.

Thus for $\alpha = 0.05$ and d.f $= 8$, the value of chi-square from the table is $\chi^2_{0.05,\,8} = 15.5073$

Diagrametically it is shown as



17.3 CHI-SQUARE AS A SAMPLING DISTRIBUTION

From a normal distribution, draw a sample of size n and determine the statistic Σz_i^2, where

$$\Sigma z_i^2 = \left(\frac{x_1 - \mu}{\sigma}\right)^2 + \left(\frac{x_2 - \mu}{\sigma}\right)^2 + \ldots + \left(\frac{x_n - \mu}{\sigma}\right)^2$$

Repeat this process for all possible samples of size n to get all values of Σz_i^2.

The sampling distribution of the statistic Σz_i^2 is a chi- square distribution with n degrees of freedom. written as $\chi^2_{(n)}$.

Note that the statistic Σz_i^2 is also denoted by $\chi^2_{(n)}$.

17.4 TEST OF GOODNESS - OF - FIT

A goodness of fit test is used to know whether or not a given set of data follows a specified probability distribution. For this purpose Karl Pearson proposed a test statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where O are observed frequencies of a set of n categories or classes and E are computed expected frequencies for these categories or classes. This test statistic approximately follows a chi-square distribution with $(n-1)$ degrees of freedom if the sample is taken from the specified distribution.

The approximation is sufficiently accurate when :

(i) There are two categories and each $E_i \geq 10$.

(ii) There are more than two categories and each $E_i \geq 5$

The testing procedure involves the following steps:

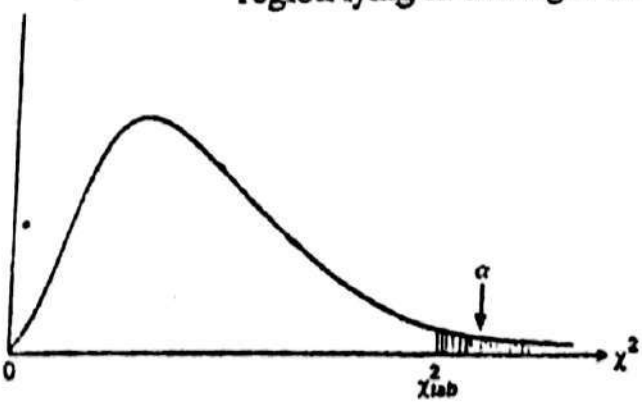
1. **Null hypothesis:** H_o : There is agreement between observed and expected frequencies.

2. **Level of significance :** $\alpha = 0.01, 0.05$ or 0.10 etc

3. **Test - statistic :** $\chi^2 = \sum \frac{(O-E)^2}{E}$; d. f. = n − 1.

4. **Critical Region :** Goodness-of-fit tests are one tailed tests with the critical region lying in the right tail.

$\chi^2_{tab} = \chi^2_{d.f.,\alpha}$



5. **Rejection Rule :** Reject H_o if $\chi^2_{cal} > \chi^2_{tab}$

6. **Conclusion:** Reject or do not reject H_o on the basis of the rejection rule.

Example: 17.4.1

A die is rolled 48 times with the following observations:

Dots on the top face (X)	1	2	3	4	5	6
Observed Frequency (O)	4	7	8	13	11	5

Test the hypothesis that the die is fair use $\alpha = 0.01$.

Solution:

In a fair die all faces are equally likely to appear, the distribution would be specified as

X	1	2	3	4	5	6
P(X)	1/6	1/6	1/6	1/6	1/6	1/6

1. Null hypothesis: H_o : The die is fair
2. Level of significance: $\alpha = 0.01$
3. Test - statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$;

   $d.f = n - 1$

Computations are shown below:

X	P(X)	O	E = N.P(X)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
1	1/6	4	8	16	2.000
2	1/6	7	8	1	0.125
3	1/6	8	8	0	0.000
4	1/6	13	8	25	3.125
5	1/6	11	8	9	1.125
6	1/6	5	8	9	1.125
Total	1	N = 48	N = 48		7.500

Same $\qquad$ X^2_{cal}

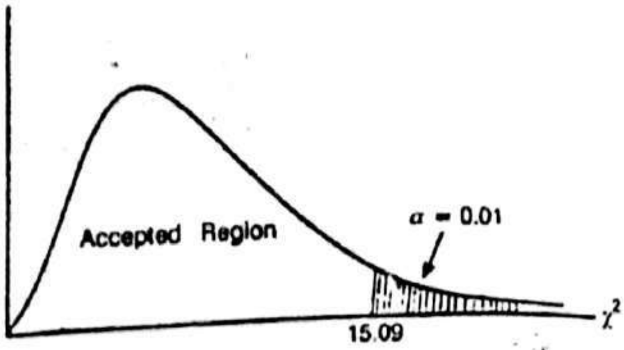
$\chi^2_{cal} = 7.5$;

4. Critical Region.

   $\alpha = 0.01$

   d.f $= 5$

   $\chi^2_{tab} = \chi_{0.01,5} = 15.09$



5. Rejection Rule :

   If $\chi^2_{cal} > \chi^2_{tab}$, reject H_o.

6. Conclusion :

$$\chi^2_{cal} = 7.5 \quad \text{and} \quad \chi^2_{tab} = 15.09$$

Since χ^2_{cal} does not exceed χ^2_{tab}, H_o can not be rejected and therefore it may be concluded that the die is fair.

Note: If the expected frequencies are less than 5, the adjacent categories (or classes) are pooled in order to make each expected frequency at least 5. The degrees of freedom will then be based upon the revised categories.

Example: 17.4.2

A company routinely purchases a certain type of bolts. The purchasing department of the company has been instructed to spread the purchase orders among suppliers A, B, C and D in the ratio of 2 : 2 : 1 : 1.

As a check, 24 purchase orders are randomly selected and suppliers A, B, C and D have received 13, 4, 4 and 3 orders respectively.

Does this indicate that th instructions are being followed:

use $\alpha = 0.05$

Solution:

1. Null hypothesis : H_o : The instructions are being followed.
2. Level of significa..ce: $\alpha = 0.05$
3. Test - statistic : $\chi^2 = \sum \frac{(O-E)^2}{E}$

Computations:

Suppliers (X)	P(X)	O	E = N.P (X)
A	2/6	13	8
B	2/6	4	8
C (Pooled)	1/6	4	4 (Pooled)
D (Pooled)	1/6	3	4 (Pooled)
Total	1	N = 24	24

Since two expected frequencies are less than 5, the adjecent categorjes are combined, giving

Suppliers (X)	P(X)	O	E	$(O - E)^2$	$\frac{(O-E)^2}{E}$
A	2/6	13	8	25	3.125
B	2/6	4	8	16	2.000
C & D	2/6	7	8	1	0.125
Total	1	24	24		5.250

$\chi^2_{cal} = 5.25$

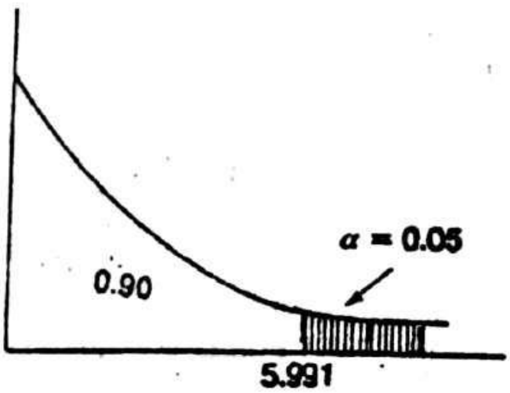
χ^2_{cal}

4. **Critical region :**

$\alpha = 0.05$

$d.f. = 3 - 1 = 2$

$\chi^2_{tab} = \chi^2_{0.05,2} = 5.991$



5. **Rejection Rule :**

If $\chi^2_{cal} > \chi_{tab}$, reject H_o

6. **Conlusion :**

$\chi^2_{cal} = 5.25$ & $\chi^2_{tab} = 5.991$

Since χ^2_{cal} is not greater than χ^2_{tab}, H_o can not be rejected. It may be concluded that the purchasing department is following the company's instructions.

Example 17.4.3

Four techniques, A, B, C and D, are currently used by business units to forecast demand for their product or service. To find out whether one technique is preferred to any other, a random sample of 200 business units were asked which technique they preferred. Their responses are shown below:

A	B	C	D
48	68	45	39

Is their sufficient evidence to indicate that there are differences in the proportions of business units preferring each technique?

use $\alpha = 0.05$

Solution :

1. Null Hypothesis : H_o: There are no differences in preferring each technique.
2. Level of significance : $\alpha = 0.05$
3. Test statistic : $\chi^2 = \sum \frac{(O-E)^2}{E}$; d.f = n − 1

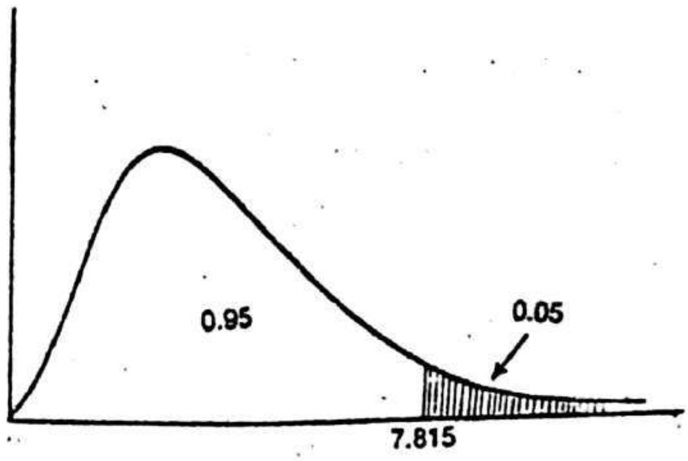
Computations:

Techniques (X)	P(X)	O	E = N.P(X)	$(O - E)^2$	$\frac{(O-E)^2}{E}$
A	0.25	48	50	4	0.08
B	0.25	68	50	324	6.48
C	0.25	45	50	25	0.50
D	0.25	39	50	121	2.42
	1.00	N = 200	200		9.48

$$\chi^2_{cal} = 9.48$$

4. **Critical Region:**

$\alpha = 0.05$



- $d.f = 4 - 1 = 3$
- $\chi^2_{tab} = \chi^2_{0.05,3} = 7.815$

5. **Rejection Rule :**

   Reject H_o if $\chi^2_{cal} > \chi^2_{tab}$

6. **Conclusion:**

   $\chi^2_{cal} = 9.48$ & $\chi^2_{tab} = 7.815$

   Since $\chi^2_{cal} > \chi^2_{tab}$, H_o is rejected.

It may be concluded that there are differences in preferring each technique.

CHAPTER 7

COUNTING TECHNIQUES

This chapter introduces some rules and definitions of counting the possible outcomes of experiments, which lays the foundation for the next chapter "Probability Theory".

An experiment is any well-defined operation or procedure that results in one of two or more possible outcomes. An outcome is a particular result of an experiment.

For example:

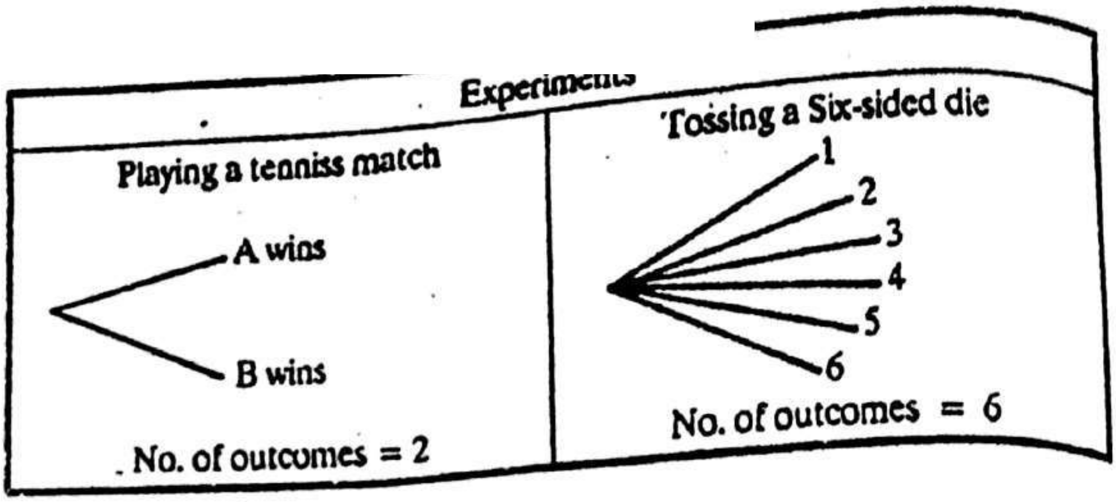
"A tennis match played between two players A and B" is an experiment. The two possible outcomes are "A wins" and "B wins".

"A single six-sided die tossed" is an experiment with six possible outcomes identified as 1, 2, 3, 4, 5 and 6.

9.1 TREE DIAGRAM

Counting the number of possible outcomes of an experiment plays a major role in probability theory. These possible outcomes can be shown by the branches of a tree-like diagram called Tree Diagram or Branch Diagram.

The tree diagrams for the above two experiments are drawn on the next page

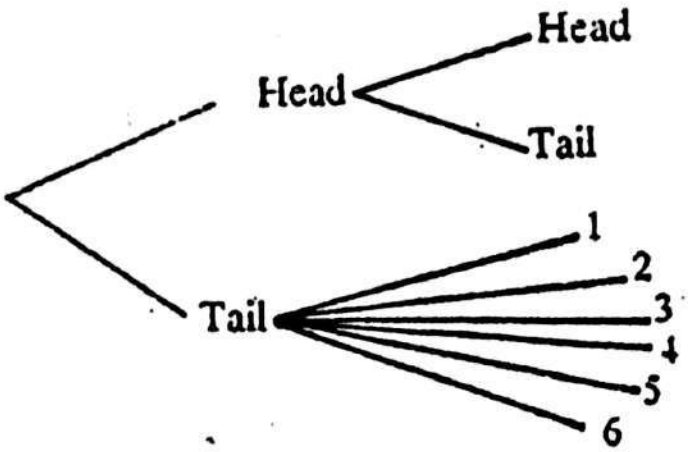


Example 9.1.1

Experiment: "A coin is tossed. If head appears, another coin is tossed otherwise a die is tossed."

Draw a tree diagram showing possible outcomes.

Solution :

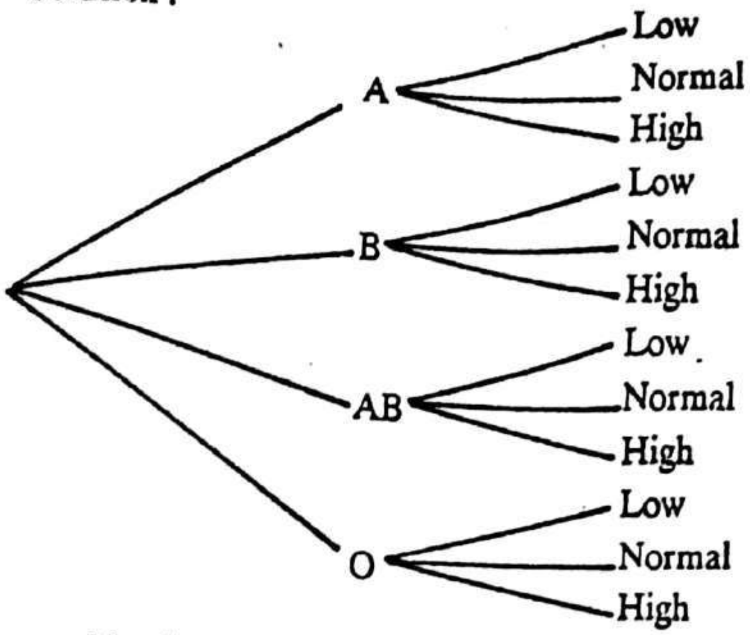


No. of outcomes = 8, these outcomes can be listed as - {HH, HT, T1, T2, T3, T4. T5,T6} where H stands for Head and T for tail.

Example 9.1.2

Experiment: "In a medical study, patients are classified according to their blood groups A, B, AB or O and also according to their blood pressures Low, Normal, or High". Draw a tree diagram to find the number of ways in which a patient can be classified.

Solution :



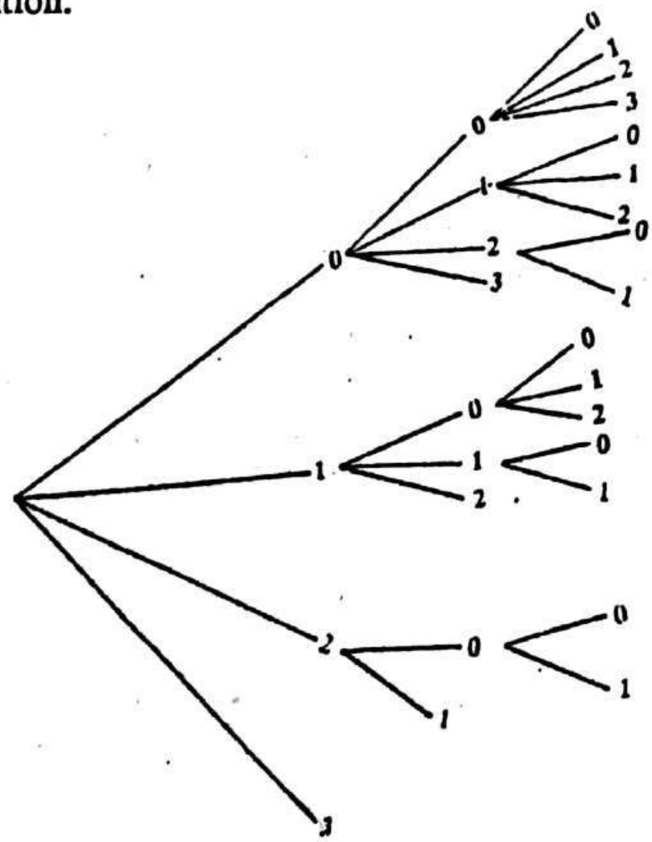
Number of ways or outcomes = 12

Example 9.1.3

Experiment: "Three law students appear in a bar examination" They can pass the examination in at most three attempts. How many of them pass the examination in each attempt is to be studied. Draw a tree diagram to show possible results in each examination.

Solution:

Let 0, 1, 2, and 3 represent the number of students that pass an examination.



No. of possibilities = 20

For experiments with large number of outcomes, the following methods are used to count the number of outcomes of the experiment without drawing a tree diagram.

1. Multiplication Rule
2. Permutations
3. Combinations

9.2 MULTIPLICATION RULE:

It is the fundamental principle of counting the number of outcomes or number of ways in which an experiment can result. The rule states that

> *If an operation of an experiment can be performed in n_1 ways and if for each of these ways another operation can be performed in n_2 ways, then the two operations can be performed together in $n_1 \cdot n_2$ ways.*

9.3 PERMUTATIONS

> *A permutation is a group of items with a certain ordered arrangement.*

Three letters A, B and C can be arranged in the following ways:

ABC, ACB, BAC, BCA, CAB, CBA.

Each arrangement is a permutation. Hence these are six different permutations of the same group of three letters.

The rules for the number of permutations under different situations are described one by one.

Situation 1:

> If (i) *All the items are distinct,*
>
> (ii) *Each item can occur only once in an arrangement. (Called "repetition not allowed" or "without replacement")*
>
> *and* (iii) *Each item can occupy any place in an arrangement.*
>
> *Then the number of Permutations of n items arranged r at a time, denoted by* ^nP_r, *is*
>
> $$^nP_r = \frac{n!}{(n-r)!} \qquad r \leq n$$

Example 9.3.1

How many three-digit numbers can be formed from the digits 1, 2, 4, 5 and 9 when each digit is used only once?

Solution:

The formula $^nP_r = \frac{n!}{(n-r)!}$ can be applied because

(i) all digits are distinct

(ii) each digit be used once only.

(iii) each digit can be placed at any of three positions.

Here $n = 5$ and $r = 3$

So, $^5P_3 = \frac{5!}{(5-3)!} = \frac{5!}{2!} = \frac{5 \times 4 \times 3 \times 2!}{2!}$

$= 60$ numbers

The same result can be obtained using the boxes of Multiplication Rule.

Hundred		Ten		Unit		
5 ways	×	4 ways	×	3 ways	=	60 numbers

Example 9.3.2

How many three-digit odd numbers can be formed from the digits 1, 2, 4, 5 and 9 when each digit is used only once?

Solution:

For the number is to be odd, the unit place of the three-digit number should be odd and hence the even digits 2 and 4 cannot occupy the unit place. Because of this restriction, general formula of permutations cannot be used. The problem is thus solved using the fundamental principle of counting.

Hundred		Ten		Unit		
3 ways	×	4 ways	×	3 ways	=	36 numbers
any of the remaining 3 digits		any digit except the odd digit at unit place		any odd digit (This place is filled first because of the restriction)		

Situation 2 :

If (i) All the items are distinct

(ii) An item can be repeated in an arrangement (Called " repetition allowed " or " with replacement ")

and (iii) Each item can occupy any place in an arrangement.

Then the number of permutations of n items arranged r at a time is

$$^nP_r = (n)^r$$

Example 9.3.3

How many three-digit numbers can be formed from the digits 1, 2, 4, 5 and 9 when the digits can be repeated.

Solution :

$$n = 5 \quad \text{and} \quad r = 3$$

Since repetition is allowed, therfore

$$^5P_3 = (5)^3 = 125 \text{ numbers}$$

Example 9.3.4

How many licence plates of three letters followed by three digits can be made if the letters and digits can be repeated.

Soluiton :

Letters = 26 & digits = 10 (including zero)

Three letters can be arranged in $(26)^3 = 17576$ ways

Three digits can be arranged in $(10)^3 = 1000$ ways.

Now the number of licence plates that can be made is $17576 \times 1000 = 17576000$ ways

If all items are not distinct, the number of permutations taken all at a time is given by the rule:

Situation 3 :

For n nondistinct items out of which n_1 are of one kind, n_2 are of another kind......., n_k are of another kind and $n_1 + n_2$ -----. $n_k = n$, the number of permutations of all n items is

$$^nP_{n_1, n_2, n_3 \ldots n_k} = \frac{n!}{(n_1)!\,(n_2)!\ldots,(n_k)!}$$

Example 9.3.5

Find the possible permutations of 7558

Solution:

$n = 4$ (4 digits 7, 5, 5, 8)

$n_1 = 1$ (One Seven)

$n_2 = 2$ (Two Fives)

$n_3 = 1$ (One Eight)

Now

$$^4P_{1,2,1} = \frac{4!}{(1)!(2)!(1)!} = \frac{4\times3\times2}{2\times1} = 12$$

This result can be verified by listing the possible arrangements which are 12 :

5578 , 5857 , 7855

5587 , 5875 , 8557

5758 , 7558 , 8575

5785 , 7585 , 8755

Example 9.3.6

How many permutations of all letters can be made from the word " COMMITTEE."

Solution:

There are nine letters in all so $n = 9$

letter M occurs twice $n_1 = 2$

letter T occurs twice $n_2 = 2$

letter E occurs twice $n_3 = 2$

The remaining letters C, O and I occur once

$\therefore \quad n_4 = n_5 = n_6 = 1$

Thus $n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 9$

Now the number of permutations is calculated as

$$^9P_{2,2,2,1,1,1} = \frac{9!}{2!\times2!\times2!\times1!\times1!\times1!}$$

$$= \frac{9\times8\times7\times6\times5\times4\times3\times2}{2\times2\times2}$$

$$= 45360$$

When the items are arranged in a circle, two arrangements are not considered different, unless corresponding items of both are preceded or followed by a different item.

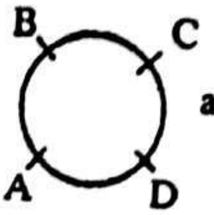
Situation 4 :

> *The number of permutations of n distinct items arranged in a circle is $(n-1)!$*

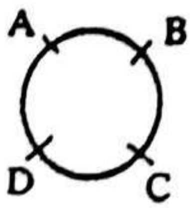
For Example

$\overline{A\ B\ C\ D}$ and $\overline{B\ C\ D\ A}$ are different permutations

but



and



are same permutaions because of the circular arrangement.

Examples 9.3.7

In how many ways can 5 different trees be planted in a circle?

Solution :

$n = 5$ trees

number of circular permutations $= (n-1)!$

$= (5-1)! = 4!$

$= 24$ ways

9.4 COMBINATIONS

> *A combination is a group of items without regard to the arrangement of items*

ABC and BCA are two different permutations but are same combinations of three letters.

> *The number of combinations of n distinct items taken r at a time, denoted by ^nC_r, is*
>
> $$^nC_r = \frac{n!}{r!\,(n-r)!}$$

Example 9.4.1

In how many ways can an instructor select five students for a group project out of a class of 12 ?

Solution:

$$n = 12$$
$$r = 5$$

Since order is not important, the number of possible groups is

$$^{12}C_5 = \frac{12!}{5!(12-5)!}$$

$$= \frac{12!}{5!\ \ 7!}$$

$$= \frac{12 \times 11 \times 10 \times 9 \times 8 \times 7!}{5 \times 4 \times 3 \times 2 \times 1 \times 7!}$$

$$= 792$$

Example 9.4.2

A box contains 7 white balls and 3 red balls. Three balls are drawn at random. In how many ways can the three balls be drawn if

(a) The colour is not considered?

(b) Two balls are white and one is red?

(c) All three balls are white

(d) At least one ball is white?

(e) All three balls are red.

Solution

(a) There are 10 balls in all. Three balls can be drawn in $^{10}C_3$ ways, where

$$^{10}C_3 = \frac{10!}{3!(10-3)!} = \frac{10!}{3!\ \ 7!} = \frac{10 \times 9 \times 8}{3 \times 2} = 120$$

(b) Two white from 7 white balls can be drawn in $^7C_2$ ways and one red from 3 red balls can be drawn in $^3C_1$ ways Then by the fundamental principle of counting, two white from 7 white AND one red from 3 red can be drawn in

$$^7C_2 \times {}^3C_1 = \frac{7!}{2!\,5!} \times \frac{3!}{1!\ \ 2!}$$

$$= 21 \times 3 = 63 \text{ ways}$$

(c) $^{7}C_{3} \times {}^{3}C_{0} = \dfrac{7!}{3!(7-3)!} \times \dfrac{3!}{0!(3-0)!}$

$= \dfrac{7!}{3! \times 4!} \times \dfrac{3!}{0! \times 3!} = \dfrac{7 \times 6 \times 5}{3 \times 2 \times 1} \times 1$

$= 35$ ways

(d) At least one white = either one white or two white or 3 white.

One white or one white and two red in $^{7}C_{1} \times {}^{3}C_{2}$ ways.

Two white or two white and one red in $^{7}C_{2} \times {}^{3}C_{1}$ ways.

Three whtie or all white and no red in $^{7}C_{3} \times {}^{3}C_{0}$ ways.

$\therefore$ At least one white

$= \left({}^{7}C_{1} \times {}^{3}C_{2} + {}^{7}C_{2} \times {}^{3}C_{1} + {}^{7}C_{3} \times {}^{3}C_{0} \right)$ ways

$= 7 \times 3 + 21 \times 3 + 35 \times 1$

$= 21 + 63 + 35 = 119$ ways

(e) $^{7}C_{0} \times {}^{3}C_{3} = \dfrac{3!}{3!\ 0!} = 1$ [Note $0! = 1$]

Many statistical principles and procedures are based on the important concept of probability.

> *A probability is a numerical measure of the likelihood (or chance) that a particular event will occur.*

The probability of any event must satisfy the following two conditions:

(i) No probability is negative ; P(Event) $\geq$ 0

(ii) No probability is greater than one ; P (Event) $\leq$ 1

There are three different approaches to assign probabilities to the events.

1. Classical or Mathematical Approach
2. Empirical or Relative Frequency Approach.
3. Subjective Approach.

10.8 CLASSICAL APPROACH OF PROBABILITY

It is the approach in which probabilities are assigned to the events before the experiment is actually performed and therefore such probabilities are also called "a priori" probabilities.

The rule to assign such probabilities is defined below:

> If (i) the possibility space of the experiment is finite and
> (ii) each outcome of the possibility space is equally likely to occur.
>
> then Probability of an event A
>
> $$= \frac{\text{number of outcomes in the event } A}{\text{number of outcomes in the possibility space } S}$$

Symbolically it is written as

$$P(A) = \frac{n(A)}{n(S)}$$

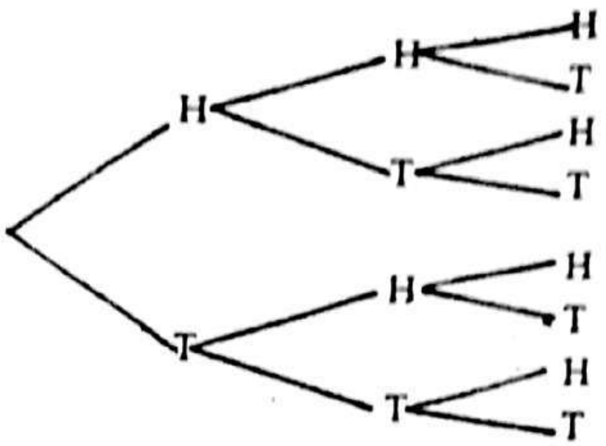
This definition is also referred to as axiomatic definition of probability.

Example 10.8.1

Three conins are tossed once. What is the probability that two heads will appear?

Solution :

A tree diagram is drawn below to find the possibility space S.



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

n (S) = 8

Event = Two Heads appear

or A = { HHT, HTH, THH }

n (A) = 3

Now probability of event A is given by

$$P(A) = \frac{n(A)}{n(S)}$$

$$= \frac{3}{8} \quad or \quad 0.375$$

Example 10.8.!

Two dice are tossed. What is the probability that the sum of the dots on the top face of both the dice is 9?

Solution:

The possibility space is

$$S = \{ (1,1) \quad (1,2) \quad (1,3) \quad (1,4) \quad (1,5) \quad (1,6)$$
$$(2,1) \quad (2,2) \quad (2,3) \quad (2,4) \quad (2,5) \quad (2,6)$$
$$(3,1) \quad (3,2) \quad (3,3) \quad (3,4) \quad (3,5) \quad (3,6)$$
$$(4,1) \quad (4,2) \quad (4,3) \quad (4,4) \quad (4,5) \quad (4,6)$$
$$(5,1) \quad (5,2) \quad (5,3) \quad (5,4) \quad (5,5) \quad (5,6)$$
$$(6,1) \quad (6,2) \quad (6,3) \quad (6,4) \quad (6,5) \quad (6,6) \}$$

$n(s) = 36$

Event = Sum of dots is 9

or $B = \{ (3,6), (4,5), (5,4), (6,3) \}$

$n(B) = 4$

$$P(B) = \frac{n(B)}{n(S)}$$

$$= \frac{4}{36} = \frac{1}{9} \text{ or } 0.1111$$

Example 10.8.3

A card is drawn from a deck of 52 cards. What is the probability that the card is a King?

Solution:

$n(S) = 52$ (52 Cards)

$n(K) = 4$ (4 Kings)

$$P(K) = \frac{n(K)}{n(S)}$$

$$= \frac{4}{52} = \frac{1}{13}$$

Example 10.8.4

A bag contains 4 Red and 6 Green balls. A ball is drawn at random from the bag. What is the probability that the ball is Red?

Solution:

$n(S) = 10$ (any of 10 balls can be drawn)

$n(R) = 4$ (4 Red balls)

$$P(R) = \frac{n(R)}{n(S)}$$

$$= \frac{4}{10} = \frac{2}{5} \text{ or } 0.4$$

Example 10.8.5

5 Cards are drawn at random from a deck of 52 playing cards. What is the probability that 2 are Kings and 3 Queens ?

Solution:

5 cards from 52 cards can be drawn in ${}^{52}C_5$ ways

$n(S) = {}^{52}C_5$

2 Kings from 4 kings can be drawn in ${}^{4}C_2$ ways and

3 Queens from 4 queens can be drawn in ${}^{4}C_3$ ways.

Therefore 2 kings and 3 Queens can be drawn in

${}^{4}C_2 \times {}^{4}C_3$ ways

$$n(A) = {}^{4}C_2 \times {}^{4}C_3$$

$$\therefore \quad P(A) = \frac{{}^{4}C_2 \times {}^{4}C_3}{{}^{52}C_5}$$

$$= \frac{6 \times 4}{2598960}$$

$$= \frac{1}{108290}$$

This method to find probability is called Combinatorial Analysis.

Another example of combinatorial analysis of probability follows:

Example 10.8.6

5 Cards are drawn at random from a deck of 52 playing cards. What is the probability that 2 cards are kings?

Solution:

Note the difference between this example and the previous example 10.8.5.

- In this example the remaining 3 cards have not been specified as Queens but may be of any denomination and colour from the remaining 48 non-king cards.

2 Kings can be drawn in ${}^{4}C_2$ ways.

3 Cards (not King) can be drawn in ${}^{48}C_3$ ways.

2 Kings and 3 other cards can be drawn in ${}^{4}C_2 \times {}^{48}C_3$ ways

$n(A) = {}^{4}C_2 \times {}^{48}C_3$

5 cards from 52 cards can be drawn in ${}^{52}C_5$ ways.

$$n(S) = {}^{52}C_5$$

$$\therefore \quad P(A) = \frac{n(A)}{n(S)}$$

$$= \frac{{}^{4}C_2 \times {}^{48}C_3}{{}^{52}C_5} = \frac{6 \times 17296}{2598960}$$

$$= \frac{2162}{54145} = 0.399$$

CHAPTER 9

SAMPLING DISTRIBUTION THEORY

14.1 POPULATION AND SAMPLE

A population is a well defined group of individuals whose characteristics are to be studied. A population containing a finite number of individuals (or Units) is called a finite population and a population with infinite number of units is called an infinite population.

A sample is a part of the population which is to be studied.

14.2 SAMPLING

The procedure by which we select or draw samples from a given population is called sampling. The aim of sampling is to get maximum information about the population from which the sample is drawn.

14.3 PROBABILITY SAMPLING

A sampling procedure in which every unit of the population has a known probability, not necessarily equal, of being included in the sample is called a probability sampling. The following sampling procedures are probability sampling.

(i) Simple Random Sampling

(ii) Stratified Random Sampling

(iii) Cluster Sampling
(iv) Systematic sampling

14.4 SIMPLE RANDOM SAMPLING

Let a finite population contains N units (called sampling units) all of which are distinguished from one another. The number of distinct samples of size n that can be drawn from the N units is given by

$$^{N}C_{n} = \frac{N\,!}{n\,!\,(N-n)\,!}$$

Then the Simple Random Sampling is a method of selecting n units out of N units such that every one of the $^{N}C_{n}$ samples has an equal chance of being chosen.

A simple random sample is drawn by one of the following devices.

(i) Tickets numbered from 1 to N for N units of the population are placed in a basket and then n units of the sample are drawn one by one.

(ii) Random Number Tables have been constructed and published to draw random samples.

(iii) Computers are used to draw random samples.

In practice a Simple Random Sample is drawn unit by unit. At any stage in the draw the process gives an equal chance of selection to all units not previously drawn. The unit drawn from the population is not replaced since this might allow the same unit to enter the sample more than once. This is described as the sampling *without replacement*.

Sampling is said to be ***with replacement*** if the sampling unit drawn from the population is returned to the population before the next unit is drawn. Sampling with replacement is done to develope some theoretical concepts.

Simple Random Sampling can be referred to as Random Sampling and the sample obtained by this procedure as Random Sample. The word "Simple" is used to differentiate it from Stratified Random sampling.

14.9 SAMPLING DISTRIBUTION

If all possible values of a statistic are considered which can occur when all possible samples are drawn from a population, a probability distribution of the statistic can be formed which is known as the sampling distribution of that statistic.

The mean of the sampling distribution of a statistic is referred to as the mean or expected value of that statistic.

The standard deviation of the sampling distribution of a statistic is called the *standard error* of that statistic.

The mean and standard error of any statistic are of great importance in the study of inferential statistics.

Some special sampling distribution are being discussed here.

14.10 SAMPLING DISTRIBUTION OF MEAN ($\bar{x}$)

From a finite population of N units with mean μ and standard deviation σ, draw all possible random samples of size n. Find the mean $\bar{x}$ of every sample. Statistic $\bar{x}$ is now a random variable. Form a probability distribution of $\bar{x}$. This distribution is called sampling distribution of mean.

The sampling distribution of mean is one of the most fundamental concepts of statistical inference and it has the following remarkable properties.

Property 1. The mean of the sampling distribution of mean is equal to the population mean. That is

$$\mu_{\bar{x}} = \mu \quad or \quad E(\bar{x}) = \mu$$

Property 2. If the sampling is done without replacement from a finite population, the standard error of mean is given by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

The fraction (N – n) / (N – 1) is called the finite population correction (f.p.c.) and $\frac{n}{N}$ is called the sampling fraction. The f.p.c.

approaches one in each of the following three cases.

(i) When the population is infinite.

(ii) When sampling fraction n/N is less than 0.05

(iii) When the sampling is with replacement.

Note : Whenever the sampling is with replacement, the population is considered infinite.

For example, consider a box with 5 balls in it. If, when a sample is drawn, the balls are replaced each time they are drawn, a sample of size $n = 10$ or $n = 100$ or whatever size is desired can be drawn. Hence the population is considered infinite.

f.p.c is necessary in theory but seldom necessary in practice because the sampling fraction $\frac{n}{N}$ is usually less than 0.05 which makes f.p.c close to one.

Property 3. When f.p.c approaches one the standard error of mean is simplified as

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

or the variance of mean as

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \quad or \quad v(\bar{x}) = \frac{\sigma^2}{n}$$

Example 14.10.1

For a population 0,4,8,12, construct sampling distribution of mean for samples of size 2 taken without replacement and find its mean and standard error.

Solution :

When sampling is done without replacement, all possible samples $= {}^NC_n = {}^4C_2 = 6$

Since every sample is equally likely to occur, the probability of each sample is $\frac{1}{6}$. The probability of each sample mean is therefore also $\frac{1}{6}$.

Sample No.	All possible samples	Sample mean $\bar{x}$	$P(\bar{x})$
1	0, 4	2	1/6
2	0, 8	4	1/6
3	0, 12	6	1/6
4	4, 8	6	1/6
5	4, 12	8	1/6
6	8, 12	10	1/6

The sampling distribution of $\bar{x}$ is tabulated below:

$\bar{x}$	$P(\bar{x})$
2	1/6
4	1/6
6	2/6
8	1/6
10	1/6

Mean $\mu_{\bar{x}}$ and standard error $\sigma_{\bar{x}}$ of this distribution are computed as.

$\bar{x}$	$P(\bar{x})$	$\bar{x}.P(\bar{x})$	$\bar{x}^2.P(\bar{x})$
2	1/6	2/6	4/6
4	1/6	4/6	16/6
6	2/6	12/6	72/6
8	1/6	8/6	64/6
10	1/6	10/6	100/6
	1	$\frac{36}{6} = 6$	$\frac{256}{6}$

$$\mu_{\bar{x}} = \Sigma\bar{x}.P(\bar{x}) = \frac{36}{6} = 6$$

$$\sigma_{\bar{x}}^2 = \Sigma\bar{x}^2.P(\bar{x}) - [\Sigma\bar{x}.P(\bar{x})]^2$$

$$= \frac{256}{6} - (6)^2$$

$$= \frac{40}{6} = \frac{20}{3}$$

$$\sigma_{\bar{x}} = \sqrt{\frac{20}{3}}$$

These results can be obtained using population mean μ and standard deviation σ.

Population mean and standard deviation are computed as

x	x^2
0	0
4	16
8	64
12	144
24	224

$$\mu = \frac{\Sigma x}{N} = \frac{24}{4} = 6$$

$$\sigma^2 = \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2$$

$$= \frac{224}{4} - (6)^2$$

$$= 20$$

$$\text{or} \quad \sigma = \sqrt{20}$$

$$\text{Now} \quad \mu_{\bar{x}} = \mu = 6$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{\sqrt{20}}{\sqrt{2}}\sqrt{\frac{4-2}{4-1}}$$

$$= \sqrt{\frac{20}{3}}$$

Example 14.10.2

For a population 0,4,8,12, construct the sampling distribution of mean for samples of size 2 taken with replacement and find its mean and standard error.

Solution :

When samples are drawn with replacement, all possible samples $= (N)^n = (4)^2 = 16$

Since 16 samples are random, every sample is equally likely to occur having probability $\frac{1}{16}$. Thus every sample mean has probability $\frac{1}{16}$.

Possible samples	sample mean ($\bar{x}$)	$P(\bar{x})$
0 , 0	0	1/16
0 , 4	2	1/16
0 , 8	4	1/16
0 , 12	6	1/16
4 , 0	2	1/16
4 , 4	4	1/16
4 , 8	6	1/16
4 , 12	8	1/16
8 , 0	4	1/16
8 , 4	6	1/16
8 , 8	8	1/16
8 , 12	10	1/16
12 , 0	6	1/16
12 , 4	8	1/16
12 , 8	10	1/16
12 , 12	12	1/16

Now the sampling distribution of $\bar{x}$ is

$\bar{x}$	$P(\bar{x})$
0	1/16
2	2/16
4	3/16
6	4/16
8	3/16
10	2/16
12	1/16
Total	1

Computation of mean and standard error.

$\bar{x}$	$P(\bar{x})$	$\bar{x}.P(\bar{x})$	$\bar{x}^2.P(\bar{x})$
0	1/16	0	0
2	2/16	4/16	8/16
4	3/16	12/16	48/16
6	4/16	24/16	144/16
8	3/16	24/16	192/16
10	2/16	20/16	200/16
12	1/16	12/16	144/16
Σ	1	96/16	736/16

$$\mu_{\bar{x}} = \Sigma \bar{x}.P(\bar{x})$$

$$= \frac{96}{16} = 6$$

$$\mu_{\bar{x}} = 6$$

$$\sigma_{\bar{x}}^2 = \Sigma \bar{x}^2 . P(\bar{x}) - \left(\Sigma \bar{x}.P(\bar{x}) \right)^2$$

$$= \frac{736}{16} - \left(\frac{96}{16}\right)^2$$

$$= 46 - 36 = 10$$

$$\sigma_{\bar{x}} = \sqrt{10}$$

These results can be obtained directly as

$$\mu_{\bar{x}} = \mu = 6$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{20}}{\sqrt{2}} = \sqrt{10}$$