Department of Statistics & O.R.

Aligarh Muslim University Aligarh

Measures of Central tendency

Measures of Location or Central Tendency or Measure of Location gives an idea about the central part of the distribution.

The main objective of Measure of Central Tendency is:

- To condense data in a single value.
- To facilitate comparison between data

Requisites of a Good Measure of Central Tendency

(i) It should be rigidly defined.

(ii) It should be simple to understand and easy to calculate.

(iii) It should be based upon all values of given data.

(iv) It should be capable of further mathematical treatment.

(v) It should have sampling stability.

(vi) It should not be unduly affected by extreme values.

There are different types of averages, each has its own advantages and disadvantages.

Arithmetic Mean

Case I: Discrete data

Arithmetic Mean (A.M.) of a set of observations is their sum divided by the number of observations

Let $X_1, X_2, \ldots, X_n$ are n observations, then

$$A.M. = \bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Example 2.1: Let there are five numbers 5, 4, 3, 7, 6, then arithmetic mean is

$$A.M. = \bar{X} = \frac{5+4+3+7+6}{5} = \frac{25}{5} = 5$$

Case II: Discrete or ungrouped frequency distribution

In case of frequency distribution $X_i \mid f_i; i = 1, 2, \ldots, n$, where f_i is the frequency of the variable X_i then

$$A.M. = \bar{X} = \frac{f_1 X_1 + f_2 X_2 + \ldots + f_n X_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N}\sum_{i=1}^{n} f_i X_i$$

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

Example 2.2: Let observations and their frequencies are given as

X: 3	5	7	8	9	10
f: 2	3	1	2	3	2

Then arithmetic mean is calculated as

$$A.M. = \bar{X} = \frac{3\times2+5\times3+7\times1+8\times2+9\times3+10\times2}{2+3+1+2+3+2} = \frac{91}{13} = 7$$

Case III: Grouped or continuous frequency distribution

In case of grouped or continuous frequency distribution, X is taken as the midpoint of the corresponding class.

Example: Distribution of marks of students in a class are given as below:

Marks:	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
No. of Students	5	3	9	2	6

Find the arithmetic mean.

Solution:

Marks	**No. of Students** (f_i)	**Mid-Point** (X_i)	$f_i X_i$
0 - 10	5	5	25
10 - 20	3	15	45
20 - 30	9	25	225
30 - 40	2	35	70
40 - 50	6	45	270
	$\sum f_i = 25$		$\sum f_i X_i = 635$

Therefore, arithmetic mean is

$$A.M. = \bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} = \frac{635}{25} = 25.4$$

Example: The following table gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail order:

No. of Order:	10 - 12	13 - 15	16 - 18	19 - 21
Frequency :	4	12	20	14

Find the arithmetic mean.

Solution:

No. of Order	**Frequency** (f_i)	**Mid-Point** (X_i)	$f_i X_i$

10 - 12	4	11	44
13 - 15	12	14	168
16 - 18	20	17	340
19 - 21	14	20	280
	$\sum f_i = 50$		$\sum f_i X_i = 832$

Therefore, arithmetic mean is

$$A.M. = \bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} = \frac{832}{50} = 16.64$$

1.1 Properties of Arithmetic Mean:

1. ***Arithmetic mean is dependent of change of origin and scale.***
2. ***Algebraic sum of the deviation of a set of values from their arithmetic mean is always zero. That is***

$$\sum_{i=1}^{n} \left(X_i - \bar{X} \right) = 0$$
$$\sum_{i=1}^{n} f_i \left(X_i - \bar{X} \right) = 0$$

where $\bar{X}$ *is the arithmetic mean.*

Example: Find the sum of the deviations of the values 3, 4, 6, 8,14 from their mean

$$\bar{X} = \frac{3+4+6+8+14}{5} = \frac{35}{5} = 7$$

$$\sum_{i=1}^{n} \left(X_i - \bar{X} \right) = (3-7) + (4-7) + (6-7) + (8-7) + (14-7) = 0$$

3. ***The sum of the square of the deviation of a set of values is minimum when taken about mean.***

Mean of combined (composite) series

Let there be k series having $n_1, n_2, \ldots, n_k$ observations respectively

Series Nos.	**1**	**2**	**3**	$\ldots$	k
	X_{11}	X_{21}	X_{31}	$\cdots$	X_{k1}
	X_{12}	X_{22}	X_{32}	$\cdots$	X_{k2}
Observations	X_{13}	X_{23}	X_{33}	$\cdots$	X_{k3}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	X_{1n_1}	X_{2n_2}	X_{3n_3}	$\cdots$	X_{kn_k}
Means	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\cdots$	$\bar{X}_k$

then the mean of combined series is given by

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} n_i\bar{X}_i}{\sum_{i=1}^{k} n_i}$$

and for $k = 2$, the above formula reduces to

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

Example: In a class average weight of 25 boys is 48 kg. and average weight of 15 girls is 40 kg. Find the average weight of students in the class.

Solution: Here

$n_1 =$ Numbers of boys in the class =25

$\bar{X}_1 =$ average weight of boys = 48 kg

$n_2 =$ Numbers of girls in the class =15

$\bar{X}_2 =$ average weight of girls = 40 kg

Therefore, the average weight of students is

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} = \frac{25 \times 48 + 15 \times 40}{25 + 15} = \frac{1800}{40} = 45\,\text{kg.}$$

Example: In a test of English average marks of Arts student is 7 and average marks of science students is 9. The mean marks of all the students are 8. Calculate the percentage(%) of Arts and Science students.

Solution: Let

$\bar{X}_1 =$ average marks of Arts students = 7

$\bar{X}_2 =$ average marks of Science students = 9

$\bar{X} =$ average marks of all students = 8

Therefore,

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

$$\Rightarrow 8 = \frac{7 \times n_1 + 9 \times n_2}{n_1 + n_2}$$

or $8n_1 + 8n_2 = 7n_1 + 9n_2$

$$\Rightarrow n_1 = n_2$$

That means the percentage of Arts and Science students are 50 each.

Weighted Average or Weighted Mean

A method of computing a kind of arithmetic mean of a set of numbers in which some elements of the set carry more importance (weight) than others.

Let there be n observations $X_1, X_2, \ldots, X_n$ and $w_1, w_2, \ldots, w_n$ are their respective weights in the set. Then weighted mean is calculated using the formula

$$W.M. = \bar{X}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Example: Grades are often computed using a weighted average.

Suppose that homework counts 10%, quizzes 20% and tests 70%.

If Mr. Zaid has a homework grade of 92, a quiz grade of 68, and a test grade of 81, then Zaid's overall grade is

$$\bar{X}_w = \frac{.10 \times 92 + .20 \times 68 + .70 \times 81}{.10 + .20 + .70} = 79.5$$

Merits of Arithmetic Mean:

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. It is based upon all values of the given data.
4. It is capable of further mathematical treatment.
5. It is not much affected by sampling fluctuations.

Demerits of Arithmetic Mean:

1. It cannot be calculated if any observation is missing.
2. It cannot be calculated for the data with open-end classes.
3. It is affected by extreme values.
4. It cannot be located graphically.
5. It may be number which is not present in the data.
6. It cannot be calculated for the data representing qualitative characteristic.

More Examples on Arithmetic Mean

Example: The sum of deviations of a certain number of observations measured from 4 is 72 and the sum of deviations of observations measured from 7 is -3. Find the number of observations and their mean.

Solution:

Let n be the required number of observations.

We have given that

$$\sum_{i=0}^{n} (X_i - 4) = 72, \text{ therefore } \sum_{i=0}^{n} X_i - 4n = 72 \qquad \text{(i)}$$

$$\text{and } \sum_{i=0}^{n} (X_i - 7) = -3, \text{ therefore } \sum_{i=0}^{n} X_i - 7n = -3 \qquad \text{(ii)}$$

Subtracting the above two equations we get, $n = 25$ and hence $\sum_{i=0}^{n} X_i = 172$.

Therefore, required mean is given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{172}{25} = 6.88$$

Example: The mean weight of 98 students is found to be 50 lbs. It is later discovered that the frequency of the class interval (30- 40) was wrongly taken as 8 instead of 10. Calculate the correct mean.

Solution:

We have given that

Incorrect mean $\left(\bar{X}\right) = 50\,Kg$ and $\sum f_i = 98$

Therefore,

Incorrect mean $\left(\bar{X}\right) = \dfrac{Incorrect \sum f_i X_i}{\sum f_i}$

$$50 = \frac{Incorrect \sum f_i X_i}{98}$$

Therefore, $Incorrect \sum f_i X_i = 98 \times 50 = 4900$

Now, $Correct \sum f_i X_i = Incorrect \sum f_i X_i - (8 \times 35) + (10 \times 35)$

$$= 4900 - 280 + 350 = 4970$$

Also, the $Correct \sum f_i = 98 + 2 = 100$

Therefore, the correct mean

$$\bar{X} = \frac{Correct \sum f_i X_i}{Correct \sum f_i} = \frac{4970}{100}$$

$$\bar{X} = 49.701\,\text{lb}$$

Example: The average marks of three batches of students having 70, 50 and 30 students respectively are 50, 55 and 45. Find the average marks of all the 150 students, taken together.

Solution:

We have given that

Batch	**I**	**II**	**III**

Average Marks	$\bar{X}_1 = 50$	$\bar{X}_2 = 55$	$\bar{X}_3 = 45$
No. of Students	$n_1 = 70$	$n_2 = 50$	$n_3 = 30$

Let X be the average marks of all 150 students taken together.

Therefore,

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3}{n_1 + n_2 + n_3} = \frac{70\times 50 + 50\times 55 + 30\times 45}{70+50+30} = \frac{7600}{150}$$

or, $\bar{X} = 50.67$ marks.

Median

Median of a distribution is the value of the variable which divides it into two equal parts, that is the value such that the number of observations above it is equal to the number of observations below it. The median is a positional average.

Applications

Median is only average to be used while dealing with qualitative data which cannot be measured quantitatively but can be arranged in ascending or descending order of magnitude.

Case I: Discrete data

Arrange all the observations in ascending or descending order.

- If the number of observations is **odd**, then

$$\text{Median} = \left(\frac{n+1}{2}\right)\text{th observation} = X_{\frac{n+1}{2}}$$

Example 2.14: Let the observations are 3, 2, 5, 1, 9, 8, 7

Then arrange these observations like

1, 2, 3, **5**, 7, 8, 9 $\longrightarrow$ Ascending order

or 9, 8, 7, **5**, 3, 2, 1 $\longrightarrow$ Descending order

Here $n = 7$, so

$$\text{Median} = \left(\frac{7+1}{2}\right)\text{th observation} = X_{\frac{7+1}{2}} = X_4 = 5\ .$$

- If the number of observations is **even**, then

$$\text{Median} = \frac{\left(\frac{n}{2}\right)\text{th obsrvation} + \left(\frac{n}{2}+1\right)\text{th obsrvation}}{2} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

Example 2.15: Let the observations are 3, 2, 5, 1, 9, 8, 7, 6

Then arrange these observations like

1, 2, 3, **5, 6**, 7, 8, 9 $\longrightarrow$ Ascending order

or 9, 8, 7, **6, 5**, 3, 2, 1 $\longrightarrow$ Descending order

Here $n = 8$, so

$$\text{Median} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} = \frac{X_4 + X_5}{2}$$

$$\text{Median} = \frac{5+6}{2} = 5.5$$

Case II: Discrete frequency distribution

Steps for calculating median

- Find $N/2$, where $N = \sum_{i=1}^{n} f_i$
- Locate cumulative frequency (*c.f.)* just greater than $N/2$
- The value of X corresponding to that *c.f.* is the median

Example 2.16: Find median for the following data

X	f	$c.f.$
1	3	3
3	2	5
4	9	14
6	6	20
9	2	22
10	5	27
12	9	36
15	4	40

Solution: Here $N = 40$

$N/2 = 20$

- Cumulative frequency (*c.f.)* just greater than 20 is 22.
- The value of X corresponding to 22 is 9.

Therefore, Median = 9

Case III: Continuous frequency distribution (grouped data)

Steps for calculating median

- Find $N/2$, where $N = \sum_{i=1}^{n} f_i$

- Locate cumulative frequency (*c.f.)* just greater than $N/2$
- The class corresponding to that *c.f.* is the median class
- Then use following formula to calculate median

$$\text{Median} = l_m + \frac{h_m}{f_m}\left(\frac{N}{2} - C_p\right),$$

where

$l_m =$ lower limit of the median class

$f_m =$ frequency of the median class

$h =$ class width

$C_p = c.f.$ of the class preceeding to the median class

Example 2.17: Find the median wage (salary) per day of workers from the following data

Wages in dollar	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
No. of Workers	3	5	20	10	2

Solution:

Wages (in dollar)	**No. of Workers** (f)	$(c.f.)$
20 - 30	3	3
30 - 40	5	8
40 - 50	20	28
50 - 60	10	39
60 - 70	2	40

Here $N = 40$, therefore

$$N/2 = 20$$

- Cumulative frequency (*c.f.)* just greater than 20 is 28
- The class corresponding to 28 is (40 - 50). This is the median class,
- Thus the

$$\text{Median} = 40 + \frac{10}{20}(20 - 8) = 46$$

Merits

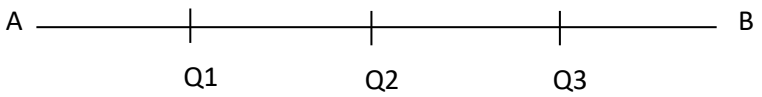
1. It is rigidly defined.
2. It is easy to calculate and understand
3. It can be located merely by inspection.

4. It is not all affected by extreme values.
5. It can be calculated for the distributions with open-end classes.

Demerits

1. In case of even number of observations median cannot be determined exactly.
2. It is not based on all observations.
3. It is not suitable for algebraic treatment.
4. As compared to mean, it is affected much by fluctuations of sampling.

Quartiles: Three points which divide the whole distribution into four equal parts are called quartiles.



Calculation of Quartiles:

Case I: Discrete data

- **The First quartile** Q_1 is the value such that 25% of the ranked data are smaller and 75% are larger.

$$Q_1 = \begin{cases} X_{\frac{n+1}{4}}, & \text{when } n \text{ is odd} \\ \frac{1}{2}\left(X_{\frac{n}{4}} + X_{\frac{n}{4}+1} \right), & \text{when } n \text{ is even} \end{cases}$$

- **The Second quartile** Q_2 is the **Median**.

- **The First quartile** Q_3 is the value such that 75% of the ranked data are smaller and 25% are larger.

$$Q_3 = \begin{cases} X_{\frac{3(n+1)}{4}}, & \text{when } n \text{ is odd} \\ \frac{1}{2}\left(X_{\frac{3n}{4}} + X_{\frac{3n}{4}+1} \right), & \text{when } n \text{ is even} \end{cases}$$

Example: Find the first quartile, second quartile (median) and third quartile of the following data of scores:

12	25	15	5	22	7	14	36	53	30	42

Solution: First, arrange the data in ascending order:

5	7	**12**	14	15	**22**	25	30	**36**	42	53
		Q_1			Q_2 (Median)			Q_3		

$$Q_1 = X_{(n+1)/4} = X_{(11+1)/4} = X_3 = 12$$

$$Q_2 = X_{2(n+1)/4} = X_{2(11+1)/4} = X_6 = 22 = \text{Median}$$

$$Q_3 = X_{3(n+1)/4} = X_{3(11+1)/4} = X_9 = 36$$

Example: Find the first quartile, median and third quartile, if 65 is added to Example 2.18.

Solution: First, arrange the data in ascending order:

5	7	12	14	15	22	25	30	36	42	53	65
			↓			↓			↓		
			Q_1			Q_2 (Median)			Q_3		

$$Q_1 = \frac{1}{2}\left(X_{\frac{n}{4}} + X_{\frac{n}{4}+1} \right) = \frac{1}{2}\left(X_3 + X_4 \right) = \frac{1}{2}(12+14) = 13$$

$$Q_2 = \frac{1}{2}\left(X_{\frac{2n}{4}} + X_{\frac{2n}{4}+1} \right) = \frac{1}{2}\left(X_6 + X_7 \right) = \frac{1}{2}(22+25) = 23.5 = \text{Median}$$

$$Q_3 = \frac{1}{2}\left(X_{\frac{3n}{4}} + X_{\frac{3n}{4}+1} \right) = \frac{1}{2}\left(X_9 + X_{10} \right) = \frac{1}{2}(36+42) = 39$$

When the subscript of X is not a whole number (whether n is odd or even), then we use the following steps

Step1: Calculate Q_1 and Q_3 using the formulae

$$Q_1 = X_{\frac{(n+1)}{4}} \text{ and } Q_3 = X_{\frac{3(n+1)}{4}}$$

Step 2: Select the ranked positions immediately below and above the number calculated.
For example, for 10 values, $Q_1 = X_{\frac{(n+1)}{4}} = X_{\frac{11}{4}} = X_{2.75}$. So, select second and third ranked values.

With these values, do the following:

- Multiply the larger ranked value by the decimal fraction of the original result (0.75 in the example).

  Multiply the smaller ranked value by 1 minus the decimal fraction of the original result (0.25 for the example, because 1-0.75 is 0.25).
- Add the two products to determine the quartile value

Special case: If the two ranked values selected be the same number, the quartile is that number and above two multiplication and addition can be skipped.

Example: Find the first and third quartiles of the following data values:

52 39 44 39 31 40 43 35 44 29

Solution: First, arrange the data in ascending order:

29	31	35	39	39	40	43	44	44	52
		$\downarrow$			$\downarrow$			$\downarrow$	
		Q_1			Q_2 (Median)			Q_3	

- Calculate $Q_1 = X_{\frac{(n+1)}{4}} = X_{\frac{(10+1)}{4}} = X_{2.75}$

  The second and third ranked values are 31 and 35.

1. Multiply the larger ranked value by the decimal fraction of the original result, we get $35 \times 0.75 = 26.25$
2. Multiply the smaller ranked value by 1 minus the decimal fraction of the original result, we get $31 \times 0.25 = 7.75$
3. Therefore final $Q_1 = 26.25 + 7.75 = 34$

- Calculate $Q_3 = X_{\frac{3(n+1)}{4}} = X_{\frac{3(10+1)}{4}} = X_{8.25}$

  The eighth and ninth ranked values are 44 and 44.

1. Multiply the larger ranked value by the decimal fraction of the original result, we get $44 \times 0.25 = 11$
2. Multiply the smaller ranked value by 1 minus the decimal fraction of the original result, we get $44 \times 0.75 = 33$
3. Therefore final $Q_3 = 11 + 33 = 44$

OR

Since eighth and ninth ranked values are same, that is 44. Therefore $Q_3 = 44$

Example: Find the first and third quartiles of the following data values:

52	39	44	39	31	40	43	35	44

Solution: First, arrange the data in ascending order:

31	35	39	39	40	43	44	44	52
		$\downarrow$		$\downarrow$			$\downarrow$	
		Q_1		Q_2 (Median			Q_3	

- Calculate $Q_1 = X_{\frac{(n+1)}{4}} = X_{\frac{(9+1)}{4}} = X_{2.5}$

  The second and third ranked values are 35 and 39.

1. Multiply the larger ranked value by the decimal fraction of the original result, we get $39 \times 0.5 = 19.50$
2. Multiply the smaller ranked value by 1 minus the decimal fraction of the original result, we get $35 \times 0.5 = 17.50$

3. Therefore final $Q_1 = 19.50 + 17.50 = 37$

- Calculate $Q_3 = X_{\frac{3(n+1)}{4}} = X_{\frac{3(9+1)}{4}} = X_{7.5}$

  The seventh and eighth ranked values are 44 and 44.

1. Multiply the larger ranked value by the decimal fraction of the original result, we get $44 \times 0.5 = 22$

2. Multiply the smaller ranked value by 1 minus the decimal fraction of the original result, we get $44 \times 0.5 = 22$

3. Therefore final $Q_3 = 22 + 22 = 44$

OR

Since eighth and ninth ranked values are same, that is 44. Therefore $Q_3 = 44$

Case II: Discrete frequency distribution (ungrouped data)

Steps for calculating quartiles

- Find $\frac{jN}{4}, j = 1, 2, 3$ where $N = \sum_{i=1}^{n} f_i$
- See cumulative frequency (*c.f.)* just greater than $\frac{jN}{4}$
- The value of X corresponding to that *c.f.* is the $Q_j, j = 1, 2, 3$

Example: Calculate all quartiles for the given data

X	f	$c.f.$
1	3	3
2	2	5
3 ←	9 ←	14
4	6	20
5 ←	2 ←	22
6	5	27
7 ←	4 ←	31
8	2	33
9	3	36
10	4	40

Solution: Here $N = 40$, thus

$$\frac{1N}{4} = 10, \frac{2N}{4} = 20 \text{ and } \frac{3N}{4} = 30$$

- Cumulative frequency (*c.f.)* just greater than 10, 20 and 30 are 14, 22, 31 respectively.

- The value of X corresponding to 14, 22, 31 are 3, 5, 7 respectively

Therefore,

$$Q_1 = 3, Q_2 = 5 \text{ and } Q_3 = 7$$

Case III: Continuous frequency distribution (grouped data)

Steps for calculating quartiles

- Find $\frac{jN}{4}, j = 1, 2, 3$ where $N = \sum_{i=1}^{n} f_i$
- See cumulative frequency (*c.f.)* just greater than $\frac{jN}{4}$
- The class corresponding to that *c.f.* is the given quartile class
- Then use following formula to calculate quartiles

$$Q_j = l_q + \frac{h}{f_q}\left(\frac{jN}{4} - C_p\right), j = 1, 2, 3,$$

where

$l_q =$ lower limit of the that quartile class

$f_q =$ frequency of the that quartile class

$h =$ class width

$C_p = c.f.$ of the class preceeding to the quartile class

Example: Calculate first and third quartiles of wage (salary) per day of workers from the following data

Wages in dollar	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80
No. of Workers	5	8	17	11	4	3

Solution:

Wages (in dollar)	**No. of Workers** (f)	$(c.f.)$
20 - 30	5	5
30 - 40	8	13
40 - 50	17	30
50 - 60	11	41
60 - 70	4	45
70 - 80	3	48

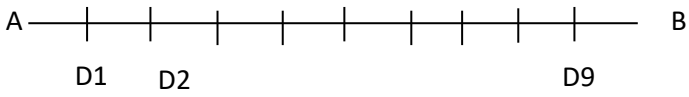
Here $N = 48$, therefore

- $\frac{1N}{4} = 12$ and $\frac{3N}{4} = 36$
- Cumulative frequency (*c.f.)* just greater than 12 is 13, so first quartile class is 30 - 40.
- Cumulative frequency (*c.f.)* just greater than 36 is 41, so third quartile class is 50 - 60.
- Thus the

$$Q_1 = 30 + \frac{10}{8}(12-5) = 38.75$$

and $$Q_3 = 50 + \frac{10}{11}(36-30) = 55.45$$

Deciles: The nine points which divide the whole distribution into ten equal parts are called deciles.



Method of Calculating Deciles

Case I: Discrete frequency distribution (ungrouped data)

Steps for calculating deciles

- Find $\frac{kN}{10}, k = 1, 2, \ldots, 9$ where $N = \sum_{i=1}^{n} f_i$
- See cumulative frequency (*c.f.)* just greater than $\frac{kN}{10}$
- The value of X corresponding to that *c.f.* is the $D_k, k = 1, 2, \ldots, 9$

Case II: Continuous frequency distribution (grouped data)

Steps for calculating quartiles

- Find $\frac{kN}{10}, k = 1, 2, \ldots, 9$ where $N = \sum_{i=1}^{n} f_i$
- See cumulative frequency (*c.f.)* just greater than $\frac{kN}{10}$
- The class corresponding to that *c.f.* is the given decile class
- Then use following formula to calculate deciles

$$D_k = l_d + \frac{h}{f_d}\left(\frac{kN}{10} - C_d\right), k = 1, 2, \ldots, 9,$$

where

l_d = lower limit of the that decile class

f_d = frequency of the that decile class

h = class width

$C_d = c.f.$ of the class preceeding to the decile class

Percentiles: The ninety-nine points which divide the whole distribution into hundred equal parts are called percentiles.

Method of Calculating Percentiles

Case I: Discrete frequency distribution (ungrouped data)

Steps for calculating deciles

- Find $\frac{lN}{100}, l = 1, 2, \ldots, 99$ where $N = \sum_{i=1}^{n} f_i$
- See cumulative frequency (*c.f.)* just greater than $\frac{lN}{100}$
- The value of X corresponding to that *c.f.* is the $P_l, l = 1, 2, \ldots, 99$

Case II: Continuous frequency distribution (grouped data)

Steps for calculating quartiles

- Find $\frac{lN}{100}, l = 1, 2, \ldots, 99$ where $N = \sum_{i=1}^{n} f_i$
- See cumulative frequency (*c.f.)* just greater than $\frac{lN}{100}$
- The class corresponding to that *c.f.* is the given percentile class
- Then use following formula to calculate deciles

$$P_l = l_p + \frac{h}{f_p}\left(\frac{lN}{100} - C_p\right), l = 1, 2, \ldots, 99,$$

where

l_p = lower limit of the that percentile class

f_p = frequency of the that percentile class

h = class width

$C_p = c.f$ .of the class preceeding to the percentile class

Example: In Example 2.23, calculate D_4, D_7, P_{30} and P_{78}.

Mode

Mode is the value which occurs most frequently in a set of observations.

Case I: Discrete frequency distribution

In case of discrete frequency distribution, mode is the value, which corresponds to maximum frequency.

Example: Find mode for the following data

X	***f***
1	4
2	9
3	16
4	25
5	22
6	15
7	7
8	3

Here maximum frequency is 25 and value corresponding to 25 is 4, therefore mode is 4.

Note: In any one or more of the following cases

- If the maximum frequency is repeated.
- If the maximum frequency occurs in the very beginning or at the end of the distribution.
- If there are irregularities in the distribution.

The value of mode is determined by the method of grouping.

Case II: Continuous frequency distribution (grouped data)

- Identify the modal class corresponding to the maximum frequency.
- Calculate the mode using the following formula

$$\text{Mode} = l + \frac{f_m - f_p}{2f_m - f_p - f_s} \times h,$$

where

l : lower limit of the modal class
h : class width
f_m : frequency of the modal class
f_p : frequency of the preceeding class
f_s : frequency of the successive class

Example: Calculate mode of the following data

Class Interval (CI)	frequency (f)
00 - 10	5
10 - 20	8
20 - 30	7
30 - 40	12
40 - 50	28
50 - 60	20
60 - 70	10
70 - 80	10

Solution: Here maximum frequency is 28, therefore modal class is 40 - 50.

Now,

$$\text{Mode} = 40 + \frac{10 \times (28 - 12)}{2 \times 28 - 12 - 20} = 40 + 6.67 = 46.67$$

Remarks:

1. In any one or more of the cases as defined for discrete case modal class is determined by the method of grouping.
2. If the method of grouping gives the modal class which does not correspond to the maximum frequency or in some cases $2f_m - f_p - f_s = 0$, in such cases mode is obtained by using formula

$$\text{Mode} = l + \frac{h(f_m - f_p)}{|2f_m - f_p| + |f_m - f_s|}$$

Merits

1. It is easy to understand and easy to calculate.
2. It is not affected by extreme values or sampling fluctuations.
3. It can be located just by inspection in many cases.
4. It is always present within data.
5. It is applicable for both quantitative and qualitative data.

Demerits

1. It is not rigidly defined.
2. It is not based upon all values of the given data.
3. It is not capable of further mathematical treatment.

Example 2.28: The marks of 36 students in an entrance test are given below:

Grades (Less than):	40	50	60	70	80	90	100
No. of Students	3	7	13	23	29	33	36

Find Mean, Median and Mode. Note that minimum mark is 30.

Solution: The given data is in cumulative frequency form. First, we shall convert this into simple frequency table.

Grades	f	cf	X	$d = \frac{X-65}{h}$	fd
30 - 40	3	3	35	–3	–9
40 - 50	4	7	45	–2	–8
50 - 60	6	13	55	–1	–6
60 - 70	10	23	65	00	00
70 - 80	6	29	75	+1	+6
80 - 90	4	33	85	+2	+8
90 - 100	3	36	95	+3	+9
	$\sum f = 36$				$\sum fd = 0$

Mean

$$\bar{X} = A + h\bar{d}$$
$$= A + h \times \frac{\sum fd}{\sum f} = 65$$

Median

Here cf just greater than $\frac{N}{2} = 18$ is 23. Thus, median class is 60 - 70.

Therefore, Median $= l_m + \frac{h_m}{f_m}\left(\frac{N}{2} - C_{pm}\right) = 60 + \frac{10}{10}(18-13) = 65$

Mode

By inspection mode lies in the class 60 - 70. Therefore

$$\text{Mode} = l_{m'} + \frac{f_{m'} - f_p}{2f_{m'} - f_p - f_s} \times h_{m'} = 60 + \frac{10-6}{20-6-6} \times 10 = 65$$

Since Mean = Median = Mode. Therefore, distribution is symmetrical.

Geometric Mean

Geometric Mean (GM) of a set of observations is the nth root of their product.

Case I: Discrete data

Let $X_1, X_2, \ldots, X_n$ are n observations, then

$$\text{GM} = \sqrt[n]{X_1 \times X_2 \times \ldots \times X_n}$$

or $$\text{GM}=\text{Antilog}\left(\frac{1}{n}\sum_{i=1}^{n}\log X_i\right)$$

For example, for a sample of two observations 4 and 9

$$GM = \sqrt{4 \times 9} = 6$$

Example: Suppose hypothetical weekly income of 9 families living in a particular locality are 70, 15, 75, 500, 8, 45, 250, 40 and 36. Compute geometric mean.

Solution:

X_i	$\log X_i$
70	1.8451
15	1.1761
75	1.8751
500	2.6990
8	0.9031
45	1.6532
250	2.3979
40	1.6021
36	1.5563
	$\sum \log X_i = 15.7079$

Therefore,

$$\text{Geometric Mean (GM)} = \text{Antilog}\left(\frac{\sum \log X_i}{n}\right) = \text{Antilog}\left(\frac{15.7079}{9}\right) = 55.63$$

Example: The annual percentage change in the sales of a popular brand of toys for five successive years was observed as 19.5, 20.8, 30.6, 28.5 and 27.2. Find the average annual percentage change in sales.

Solution:

Years	% Change	X_i	$\log X_i$
1	19.5	119.5	2.07740
2	20.8	120.8	2.08202
3	30.6	130.6	2.11591
4	28.5	128.5	2.10891

5	27.2	127.2	2.10451
			$\sum \log X_i = 10.48875$

Therefore,

Geometric Mean (GM) $= \text{Antilog}\left(\frac{\sum \log X_i}{n}\right) = \text{Antilog}\left(\frac{10.48875}{5}\right) = 125.26$

It means that sales have increased by (125.26 - 100.00) = 25.26 %

Case II: Discrete frequency distribution

In case of frequency distribution $X_i \mid f_i; i = 1, 2, \ldots, n$, where f_i is the frequency of the variable X_i, the geometric mean is calculated by using formula

$$GM = \sqrt[N]{X_1^{f_1} \times X_2^{f_2} \times \ldots \times X_n^{f_n}}$$

$$\log GM = \frac{1}{N} \sum_{i=1}^{n} f_i \log X_i$$

or $$\text{GM=Antilog}\left(\frac{1}{N} \sum_{i=1}^{n} f_i \log X_i\right),$$

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

Note: For computation of weighted geometric mean f_i is replaced with the weight of ith observation w_i.

Example: Suppose there are five observations 5, 12, 20, 50, 90 and their corresponding frequency in the data set are 15, 10, 7, 4 and 2. Compute geometric mean the given data set.

Solution:

X_i	f_i	$\log X_i$	$f_i \log X_i$
5	16	0.6990	11.1840
12	11	1.0792	11.8712
20	7	1.3010	09.1070
50	4	1.6990	06.7960
90	2	1.9031	03.8062
	$\sum f_i = 40$	$\sum \log f_i = 6.6813$	$\sum f_i \log X_i = 42.7640$

$$GM = \text{Antilog}\left(\frac{\sum f_i \log x_i}{\sum f_i}\right) = \text{Antilog}\left(\frac{42.7640}{40}\right)$$

$$GM = \text{Antilog}(1.0691) = 11.78$$

Example 2.32: Obtain geometric mean of the following distribution

Class Intervals	**10 - 20**	**20 - 30**	**30 - 40**	**40 - 50**	**50 - 60**
Frequencies	5	8	12	10	5

Solution:

Class Interval	X_i	f_i	$\log X_i$	$f_i \log X_i$
10-20	15	5	1.17609	5.8804
20-30	25	8	1.39794	11.1835
30-40	35	12	1.54407	18.5288
40-50	45	10	1.65321	16.5321
50-60	55	5	1.74036	08.7018
		$\sum f_i = 40$		$\sum f_i \log X_i = 60.8266$

$$GM = \text{Antilog}\left(\frac{\sum f_i \log x_i}{\sum f_i}\right) = \text{Antilog}\left(\frac{60.8266}{40}\right)$$

$$GM = \text{Antilog}(1.5206) = 33.16$$

Computation of Average Population Growth Rate

Example: If the population of a city has recorded a growth rate of 20% in the first decade, 30% in the second decade, and 40% in the third decade. Obtain average population growth rate.

Decade	**Population Growth Rate**	**Population at the End of Decade**	$\log X_i$
First	20%	120	2.0792
Second	30%	130	2.1139
Third	40%	140	2.1461
Total			$\sum f_i \log X_i = 6.3392$

Therefore

$$GM = \text{Antilog}\left(\frac{1}{n}\sum \log x_i\right) = \text{Antilog}\left(\frac{6.3392}{3}\right)$$

$$GM = \text{Antilog}(2.1131) = 129.7$$

It means that the city population over the three-decade period has increased at an average rate of (129.7-100) = 29.7 per cent per decade

Harmonic Mean

Harmonic mean of the given observations is defined as the reciprocal of the arithmetic mean of the reciprocals of the given set of observations.

If $X_1, X_2, ..., X_n$ are n sample observations, with their reciprocals $\frac{1}{X_1}, \frac{1}{X_2}, ..., \frac{1}{X_n}$, then the Harmonic Mean (HM) is given by

$$HM = \frac{1}{\frac{1}{n}\left(\frac{1}{X_1}, \frac{1}{X_2}, ..., \frac{1}{X_n}\right)} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}.$$

In the case of ungrouped frequency distribution

$$HM = \frac{1}{\frac{1}{N}\left(\frac{f_1}{X_1}, \frac{f_2}{X_2}, ..., \frac{f_n}{X_n}\right)} = \frac{N}{\sum_{i=1}^{n} \frac{f_i}{X_i}}, N = \sum_{i=1}^{n} f_i$$

In the case of grouped frequency distribution $X_i, i = 1, 2, ..., n$ are the mid points of the class intervals.

Note: For computation of weighted harmonic mean f_i is replaced with the weight of ith observation w_i.

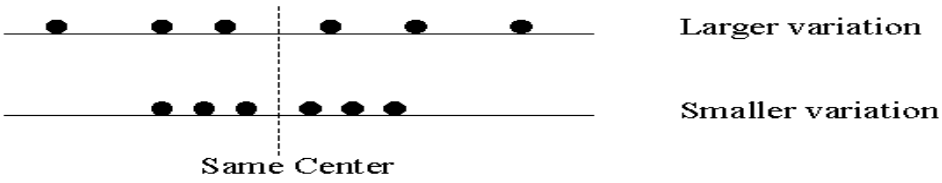
Department of Statistics & O.R.

Aligarh Muslim University Aligarh

Measure of Dispersion

Measures of Dispersion

Measure of Variability or Measure of Dispersion indicates how spread out the data around the mean.



The following are the measure of dispersions:

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

Characteristic for ideal measure of dispersion

(i) It should be rigidly defined.

(ii) It should be simple to understand and easy to calculate.

(iii) It should be based upon all values of given data.

(iv) It should be capable of further mathematical treatment.

(v) It should have sampling stability.

1. Range:

The range is the **difference between the largest and smallest values** of the distribution**.**

Let $X_1, X_2, \dots, X_n$ are the set of n observations, then

$$\text{Range} = \max(X_1, X_2, \dots, X_n) - \min(X_1, X_2, \dots, X_n)$$

Example 1: Let there are five numbers 25, 34, 13, 27, 36, then

$$\text{Range} = 36 - 13 = 23$$

- ❖ One of the simplest measures of variability to calculate.
- ❖ Depends only on extreme values and provides no information about how the remaining data is distributed.

2. Quartile Deviation

Quartile Deviation or Semi Inter Quartile Range (Q) is given by

$$Q = \frac{1}{2}(Q_3 - Q_1)$$

where Q_1 and Q_3 are the first and third quartiles of the distribution.

Example 2: Suppose there are 12 observations

12, 25, 15, 5, 22, 7, 14, 36, 53, 30, 42, 65

In order to calculate quartiles we sort the observations,

5, 7, 12, 14, 15, 22, 25, 30, 36, 42, 53, 65

Here $n = 12$ (even)

$$Q_1 = \frac{1}{2}\left(X_{\frac{n}{4}} + X_{\frac{n}{4}+1}\right) = \frac{1}{2}\left(X_{\frac{12}{4}} + X_{\frac{12}{4}+1}\right)$$

$$= \frac{1}{2}(X_3 + X_4) = \frac{1}{2}(12 + 14) = 13$$

$$Q_3 = \frac{1}{2}\left(X_{\frac{3n}{4}} + X_{\frac{3n}{4}+1}\right) = \frac{1}{2}\left(X_{\frac{3\times 12}{4}} + X_{\frac{3\times 12}{4}+1}\right)$$

$$= \frac{1}{2}(X_9 + X_{10}) = \frac{1}{2}(36 + 42) = 39$$

Therefore Quartile Deviation is given by

$$QD = \frac{1}{2}(39 - 13) = 13$$

Similarly for **discrete frequency distribution** and **grouped frequency distribution** Quartile Deviation can be calculated after getting the value of Q_1 and Q_3.

- It is better than range, as it uses 50% observations. Since it ignores other 50% observations, so it cannot be regarded as an ideal measure.

3. Mean Deviation

Case I: Discrete data

Let $X_1, X_2, \dots, X_n$ are n observations, then

$$\text{Mean Deviation (M. D.)} = \frac{1}{n}\sum_{i=1}^{n} |X_i - A|$$

where A is generally taken as mean, median and mode.

Example 3: Find mean deviation from mean for the given data 5, 4, 3, 7, 6

Solution: Here

$$A.M. = \bar{X} = \frac{5 + 4 + 3 + 7 + 6}{5} = \frac{25}{5} = 5$$

So $A = \bar{X} = 5$

Therefore mean deviation from mean is given by

$$\text{M. D.} = \frac{|5 - 5| + |4 - 5| + |3 - 5| + |7 - 5| + |6 - 5|}{5} = \frac{6}{5} = 1.2$$

Case II: Discrete frequency distribution

In case of frequency distribution $X_i|f_i; i = 1, 2, \dots, n$, where f_i is the frequency of the variable X_i, then

$$M.D. = \frac{1}{N}\sum_{i=1}^{n} f_i|X_i - A|$$

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

Example 4: Find mean deviation from mean for the given data

X:	3	5	7	8	9	10
f:	2	3	1	2	3	2

Solution:

Hare Arithmetic Mean is 7. Therefore $A = \bar{X} = 7$

X	f	$\lvert X - \bar{X}\rvert$	$f\lvert X - \bar{X}\rvert$
3	2	4	8
5	3	2	6
7	1	0	0
8	2	1	2
9	3	2	6
10	2	3	6
	$\sum f$ =13		$\sum f\lvert X - \bar{X}\rvert$ =28

Therefore

$$M.D.(\bar{X}) = \frac{1}{N}\sum_{i=1}^{n} f_i|X_i - \bar{X}| = \frac{1}{13} \times 28 = 2.15$$

Case III: Grouped or continuous frequency distribution

In case of grouped or continuous frequency distribution, X is taken as the midpoint of the corresponding class.

Example 5: Distribution of marks of students in a class are given as below:

Marks:	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
No. of Students	5	3	9	2	6

Find the mean deviation from the mean.

Solution:

Marks	No. of Students (f_i)	Mid Point (X_i)	f_iX_i	$\lvert X_i - \bar{X}\rvert$	$f_i\lvert X_i - \bar{X}\rvert$
0 - 10	5	5	25	20	100
10 - 20	4	15	60	10	40
20 - 30	8	25	200	0	0
30 - 40	2	35	70	10	20
40 - 50	6	45	270	20	120
	$\sum f_i$ =25		$\sum f_iX_i$ = 625		$\sum f_i\lvert X_i - \bar{X}\rvert$ =280

Here, arithmetic mean is

$$A.M. = \bar{X} = \frac{\sum_{i=1}^{n} f_iX_i}{\sum_{i=1}^{n} f_i} = \frac{625}{25} = 25$$

Therefore mean deviation from mean is

$$\text{M.D.}(\bar{X}) = \frac{1}{N}\sum_{i=1}^{n} f_i|X_i - \bar{X}| = \frac{1}{25} \times 280 = 11.2$$

- It is based on all the observations, it is better measure of dispersion than above two. But step of ignoring negative sign of the deviation $(X_i - A)$ creates artificiality and it is not capable of further mathematical treatment.

4. Standard Deviation

Case I: Discrete data

Let $X_1, X_2, \dots, X_n$ are n observations, then Standard Deviation generally denoted by σ is given by

$$\text{Standard Deviation (S.D.)} = \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where $\bar{X}$ is the arithmetic mean.

The square of Standard Deviation is called **Variance**. Therefore the formula for calculating Variance is

$$\text{Variance} = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

The above formulation can also be written as

$$\text{Variance} = \sigma^2 = \frac{1}{n}\{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\}$$

Example 6: Find Standard Deviation and Variance from the given data 5, 4, 3, 7, 6

Solution: Here

$$A.M. = \bar{X} = \frac{5+4+3+7+6}{5} = \frac{25}{5} = 5$$

Therefore standard deviation is given by

$$\text{S.D.} = \sigma = \sqrt{\frac{(5-5)^2 + (4-5)^2 + (3-5)^2 + (7-5)^2 + (6-5)^2}{5}} = \sqrt{\frac{10}{5}} = 1.41$$

Hence the variance is

$$\text{Variance} = \sigma^2 = 2$$

Case II: Discrete frequency distribution

In case of frequency distribution $X_i|f_i; i = 1, 2, \dots, n$, where f_i is the frequency of the variable X_i, then standard deviation is

$$S.D. = \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(X_i - \bar{X})^2}$$

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

and

$$\text{Variance} = \sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(X_i - \bar{X})^2$$

The above formulation can also be written as

$$\text{Variance} = \sigma^2 = \frac{1}{N}\{\sum_{i=1}^{n} f_i X_i^2 - N\bar{X}^2\}$$

Example 7: Find Standard Deviation for the given data

X:	3	5	7	8	9	10
f:	2	3	1	2	3	2

Solution:

Hare Arithmetic Mean is $\bar{X} = 7$.

X	f	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
3	2	16	32
5	3	4	12
7	1	0	0
8	2	1	2
9	3	4	12
10	2	9	18
	$\sum f$ =13		$\sum f(X-\bar{X})^2$ =76

Therefore

$$S.D. = \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(X_i - \bar{X})^2} = \sqrt{\frac{1}{13} \times 76} = \sqrt{5.85} = 2.42$$

Case III: Grouped or continuous frequency distribution

In case of grouped or continuous frequency distribution, X is taken as the midpoint of the corresponding class.

Example 8: Distribution of marks of students in a class is given as below:

Marks:	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
No. of Students	5	3	9	2	6

Find the standard deviation and hence variance.

Solution:

Marks	No. of Students (f_i)	Mid Point (X_i)	f_iX_i	$(X_i-\bar{X})^2$	$f_i(X_i-\bar{X})^2$
0 - 10	5	5	25	400	2000
10 - 20	4	15	60	100	400
20 - 30	8	25	200	0	0
30 - 40	2	35	70	100	200
40 - 50	6	45	270	400	2400
	$\sum f_i$ =25		$\sum f_iX_i = 625$		$\sum f_i(X_i-\bar{X})^2$ =5000

Here, arithmetic mean is

$$A.M. = \bar{X} = \frac{\sum_{i=1}^{n} f_iX_i}{\sum_{i=1}^{n} f_i} = \frac{625}{25} = 25$$

Therefore

$$S.D. = \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(X_i - \bar{X})^2} = \sqrt{\frac{1}{25} \times 5000} = \sqrt{200} = 14.14$$

and

$$\text{Variance} = \sigma^2 = 200$$

- It is based on all the observations; it is better measure of dispersion than mean deviation.
- The artificiality created in the case of mean deviation is removed by squaring the $(X_i - \bar{X})$.
- It is capable of further mathematical treatment.
- **Some Useful Results:**

Let X is any variable and K be any constant. If we denote Mean by M and Variance by V, then

For Mean

1. $\mathrm{M}(K) = K$
2. $\mathrm{M}(X \mp K) = M(X) \mp K$
3. $\mathrm{M}(KX) = K \times M(X)$
4. $\mathrm{M}(KX \mp C) = K \times \mathrm{M}(X) \mp C$,

where C is any other constant

For Variance and SD

1. $\mathrm{V}(K) = 0$
2. $\mathrm{V}(X \mp K) = \mathrm{V}(X)$
3. $\mathrm{V}(KX) = K^2 \times \mathrm{V}(X)$
4. $\mathrm{SD}(KX) = K \times \mathrm{SD}(X)$

Department of Statistics & O.R.

Aligarh Muslim University Aligarh

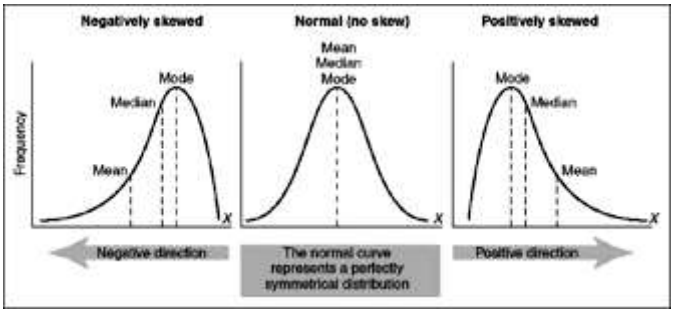
Measure of Skewness and Kurtosis

Measure of Skewness

The term skewness refers to lack of symmetry or departure from symmetry. When a distribution is not symmetrical (or is asymmetrical), it is called skewed distribution. The measures of skewness indicate the difference between the manners in which the observations are distributed in particular distribution comparedwith a symmetrical (or normal) distribution. The concept of skewness gains importance from the fact that statistical theory is often based upon the assumption of the normal distribution.

Nature of Skewness: Skewness can be positive or negative or zero.

- When the values of mean, median and mode are equal, there is no skewness.
- When mean > median > mode, skewness will be positive.
- When mean < median < mode, skewness will be negative.



There is different measure of skewness, which are discussed below.

1. **Karl Pearson's Coefficient of Skewness**

   This method is most frequently used for measuring skewness. The formula for measuring coefficient based on mean, mode and standard deviation is given by

$$sk_P = \frac{Mean - Mode}{SD},$$

   where sk_P denotes the Pearsonian coefficient of skewness.

   This formula can be used for fairly symmetric data. However, if the data is moderately skewed or mode is ill defined, then the above formula can be modified by using the empirical relation as,

$$sk_P = \frac{3(Mean - Median)}{SD}$$

   - If $sk_P = 0$, the symmetrical distribution.
   - If mean is greater than mode, implies $sk_p > 0$, then positive skewness.
   - If mean is less than mode, implies $sk_p < 0$, then negative skewness.

   The Karl Pearson's coefficient of skewness lies between -3 to $+3$.

2. **Bowley's Coefficient of Skewness**

   This method is based on quartiles. The formula for calculating coefficient of skewness is

$$sk_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1}$$

- If $sk_B = 0$, then it is symmetrical distribution.
- If $sk_B > 0$, and then positively skewed distribution.
- If $sk_B < 0$, and then negatively skewed distribution.

The Bowley's coefficient of skewness is used for moderately skewed distribution and distribution having open end class. The Bowley's coefficient of skewness lies between $+1$ to -1.

Example 5: The IQ scores of 50 students of a class are given blow

IQ Score:	50 - 60	60 - 70	70 - 80	80 - 90	90 - 100
No. of Std:	5	8	16	12	9

Calculate

(i) Karl Pearson coefficient of skewness based on median and mode.
(ii) Bowley's coefficient of skewness.

Solution: To calculate all above measures, first we shall calculate mean, median, mode, standard deviation, first quartile and third quartile.

IQ Score	No. of Std. (f)	cf	X	$d = \frac{X-A}{h}$	fd	fd^2
50 - 60	5	5	55	-2	-10	20
60 - 70	8	13	65	-1	-8	8
70 - 80	16	29	75	0	0	0
80 - 90	12	41	85	1	12	12
90 - 100	9	50	95	2	18	36
	$\sum f = 50$				$\sum fd = 12$	$\sum fd^2 = 76$

$$\bar{X} = A + h\bar{d} = 75 + 10 \times 0.24 = 77.4$$

$$\sigma_d^2 = \frac{1}{N}\sum fd^2 - \bar{d}^2 = 1.52 - 0.0576 = 1.46$$

$$\sigma_x^2 = h^2\sigma_d^2 = 100 \times 1.46 = 146$$

$$\sigma_x = 12.08$$

$$Q_1 = l_{q_1} + \frac{h}{f_{q_1}}\left(\frac{N}{4} - C_{pq_1}\right) = 60 + \frac{10}{8}(12.5 - 5) = 69.38$$

$$Q_2 = l_{q_2} + \frac{h}{f_{q_2}}\left(\frac{N}{2} - C_{pq_2}\right) = 70 + \frac{10}{16}(25 - 13) = 83.75$$

$$Q_3 = l_{q_3} + \frac{h}{f_{q_3}}\left(\frac{N}{4} - C_{pq_3}\right) = 80 + \frac{10}{12}(37.5 - 29) = 87.06$$

(i) The Karl Pearson coefficient of skewness based on mean, mode and standard deviation is given by

$$sk_P = \frac{\bar{X} - Mode}{\sigma_x} = \frac{77.4 - 76.67}{12.08} = 0.06$$

The Karl Pearson coefficient of skewness based on mean, median and standard deviation is given by

$$sk_P = \frac{3(\bar{X} - Md)}{\sigma_x} = \frac{3(77.4 - 83.75)}{12.08} = -1.58$$

(ii) The Bowley's coefficient of skewness is

$$sk_B = \frac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1} = \frac{69.38 - 2 \times 83.75 + 87.06}{87.06 - 69.38} = -0.6256$$

3. Measure of Skewness based on Moments

Karl Pearson defined the following four coefficients, based upon the first four central moments. The coefficients β_1 and γ_1 are used to measure the skewness. These coefficients are defined as below:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \quad \gamma_1 = \pm\sqrt{\beta_1} = \frac{\mu_3}{\sigma^3}$$

For a symmetrical distribution $\beta_1 = 0$, but it does not tell the direction of skewness, that is positive or negative. Because $\mu_3^2 > 0$ and variance μ_2 is always positive. This drawback is removed by calculating γ_1 coefficient. Thus, the sign of skewness would depend upon the value of μ_3 whether it is positive or negative.

- If $\gamma_1 > 0$, then positively skewed distribution
- If $\gamma_1 < 0$, then negatively skewed distribution
- If $\beta_1 = 0$, then symmetrical distribution

Measure of Kurtosis

The relative flatness of the top is called kurtosis or convexity of curve. The coefficients β_2 and γ_2 are used to measure the kurtosis. These coefficients are defined as below:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, \quad \gamma_2 = \beta_2 - 3$$

The quantity $\gamma_2 = \beta_2 - 3$ is called excess of kurtosis.

- If $\gamma_2 < 0$, then curve is called Platykurtic.
- If $\gamma_2 = 0$, then curve is called Mesokustic.
- If $\gamma_2 > 0$, then curve is called Leptokurtic.

Example : Example 2: Wages of workers are given in the following table:

Weekly wages	**10-12**	**12-14**	**14-16**	**16-18**	**18-20**	**20-22**	**22-24**
No. of workers	1	3	7	20	12	4	3

calculate $\beta_1, \gamma_1, \beta_2$ and γ_2 coefficients. Also study the nature of the distribution.

Solution: We have calculated the following first four central moments as

$$\mu_1 = 0,\ \mu_2 = 1.8896, \mu_3 = 7.491 \text{ and } \mu_4 = 126,5874$$

Therefore

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(7.491)^2}{(1.8896)^3} = \frac{56.11508}{6.7469} = 8.317$$

and $$\gamma_1 = \pm\sqrt{\beta_1} = \frac{\mu_3}{\sigma^3} = \frac{7.491}{(1.37)^3} = 2.91$$

Since $\gamma_1 > 0$, therefore distribution is positively skewed.

Further,

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{126.5874}{(1.8896)^2} = 35.4527$$

and $$\gamma_2 = \beta_2 - 3 = 32.4527$$

Since $\gamma_2 > 0$, therefore distribution is leptokurtic.

Practice Exercises

1. The first three moments of a distribution about the value 2 of the variable are 1, 16 and -40. Show that the mean is 3, the variance is 15 and $\mu_3 = -86$. Also show that the first three moments about $x = 0$ are 3, 24 and 76.

2. The first four moments of a distribution about the value 5 of the variable are 2, 20, 40 and 50. Show that the mean is 7, variance 16, $\mu_3 = -64, \mu_4 = 162, \beta_1 = 1$ and $\beta_2 = 0.63$.

3. For a certain distribution, the mean is 10, variance is 16, γ_1 is $+1$ and β_2 is 4. Find the first four moments about the origin.

4. Calculate the first four moments about the mean for the following data. Also calculate β_1 and β_2.

x:	1	2	3	4	5	6	7	8	9
f:	1	6	13	25	30	22	9	5	2

5. In a certain distribution, the first four moments about the point 4 are -1.5, 17, -30 and 108. Calculate β_1 and β_2 and state whether the distribution is leptokurtic or platykurtic.

6. Show that for discrete distribution

   (i) $\beta_2 \geq 1$

   (ii) $\beta_2 \geq \beta_1$.

Department of Statistics & O.R.

Aligarh Muslim University Aligarh

Probability Theory

In our daily life, we often used phrases such as 'It may rain today', or 'India may win the match' or ' I may be selected for this post.' These phrases involve an element of uncertainty. How can we measure this uncertainty? A measure of this uncertainty is provided by a branch of Mathematics, called the theory of probability.

Experiment: A set of action taken in a certain order with certain objective is known as experiment.

Example: Tossing of a coin or throwing of a die.

Experiment can be divided according to the nature of their outcome, which is as below:

- Deterministic experiment
- Non-deterministic experiment

1. **Deterministic experiment**

   An experiment whose outcome will be same whenever it is repeated under the identical condition is known as deterministic experiment.

   Example: Mixing of certain chemicals in a same substance.
   Mining at a constant speed.
   Time taken to travel a certain distance.

2. **Non-deterministic experiment or random experiment**

   An experiment which, though repeated under essentially identical conditions, does not give unique result but may result in any one of the several possible outcomes is called non-deterministic experiment or **random experiment**.

   In other words, any probabilistic situation is referred as random experiment.

   Example: Tossing of anunbiased coin results either Head or Tail coming up.
   In throwing a die any one of the six faces may come uppermost.

Sample Space:

Set of all possible outcomes of a random experiment is called sample space. It is generally denoted by S. A sample space is said to be finite/infinite if the number of elements in S are finite/infinite.

Example 1.1: If a coin is tossed, then set of all possible outcome is {H, T}. Therefore, sample space S = {H, T}

Example 1.2: If a die is thrown and we are interested in the number shown at top. Then set of all possible outcomes is {1, 2, 3, 4, 5, 6}. Therefore, sample space S = {1, 2, 3, 4, 5, 6}

Example 1.3: An experiment consists of flipping a coin and then flipping it a second time if a head occurs. If a tail occurs on the first flip, then a die is tossed once. Then set of all possible outcomes that denotes the sample space is

$$S = \{HH, HT, T1, T2, T3, T4, T5, T6\}$$

Event

The point of a sample space is called an event. An event is a subset of the sample space.

- $\phi \subseteq S$ is an event of a every sample space and is called impossible event.
- $S \subseteq S$ is an event and is called sure event.

In the example 1.1, H and T are the events.

In the example 1.2, following are the events:

E_1= getting an even number ={2,4,6}$\subseteq S$

E_2 = getting a number less than 4={1,2,3}$\subseteq S$

E_3 = getting 1 or 3={1,3}$\subseteq S$

E_4 = getting an odd number={1,3,5}$\subseteq S$

E_5 = getting a negative number={ }=$\phi \subseteq S$

E_6 = getting a number less than 10 ={1,2,3,4,5,6} = $S \subseteq S$

In the example 1.3, getting HH, T2 or T5 are the events.

Example 1.4: From manufacturing process 3 items are taken; each item is inspected and classified as defective (*D*) or non-defective (*N*).

This experiment has 8 possible outcomes. Thus, sample space can be defined as

$$S = \{DDD, DDN, DND, DNN, NDD, NDN, NND, NNN\}$$

The following events can be in this sample space

A={at least 2 defectives}= {DDD,DDN,DND,NDD}$\subseteq S$

B={at most one defective}= {DNN,NDN,NND,NNN}$\subseteq S$

C={3 defectives}={DDD}$\subseteq S$

Exhaustive Events:

The total number of possible outcomes in any trial is known as exhaustive events or exhaustive cases. In other words, the total points of a sample space is called exhaustive events.

In above example 1.1, exhaustive number of cases = 2

In above example 1.2, exhaustive number of cases = 6

In above example 1.3, exhaustive number of cases = 8

Favourable Events or Cases:

The number of cases belonging to the happening of any given event in a trial are favourable events. In other words, those points of the sample space which belong to the happening of any given event.

Examples

- In throwing of a die

  E_1= getting an even number ={2,4,6}$\subseteq S$

  E_2 = getting a number less than 4={1,2,3}$\subseteq S$

  are favourable events.
- In throwing of two dice, the number of cases favourable to getting a sum 5 is (1, 4), (4, 1), (2, 3), (3, 2), that is 4.
- In drawing a card from a pack of cards, the number of cases favourable to drawing of an ace is 4, for drawing a spade is 13 and for drawing a red card is 26.

Mutually Exclusive Events

Events are said to be mutually exclusive or incompatible, if happening of any one of them prevents the happening of all others. That is no two or more of them can happen simultaneously in the same trail. For example, in tossing of unbiased coin the events Head and Tail are mutually exclusive.In throwing of die, all the six faces are mutually exclusive.

Equally Likely Cases or Events:

Cases are said to be equally likely when we have no reason to believe that one is more likely to occur than the other. For example, in throwing of an unbiased coin Head and Tail are equally likely events and in throwing of a die, all the six faces are equally likely to come.

Independent Events

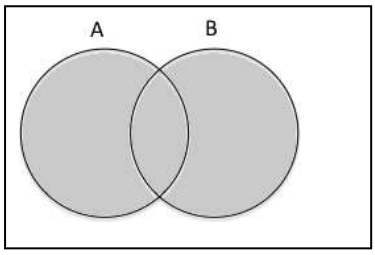
Several events are said to be independent if the happening (or non-happening) of an event is not affected by the supplementary knowledge concerning the occurrence of any number of the remaining events. For example, in throwing an unbiased coin the event of getting Head in the first toss is independent of getting a Head in the second, third and subsequent throws.

Operations on Events

Let A and B are two events defined over the sample space S of the random experiment, then following can be seen

1. **Union:** The Union of two given events A and B, denoted by $A \cup B$, is defined as a set consisting of all those points which belong to either **A or B or both**.

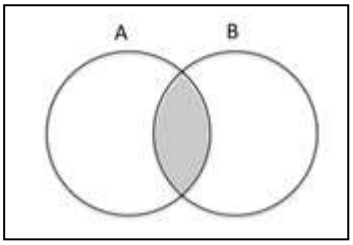
   i.e.$A \cup B = \{x: x \in A \text{ or } x \in B\}$



   For example: If $A = \{1, 2, 3, 7, 8\}$and $B = \{2, 3, 11\}$, then $A \cup B = \{1, 2, 3, 7, 8, 11\}$.

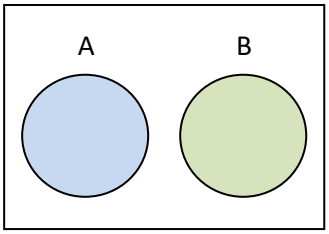
2. **Intersection:** The Intersection of two given events A and B, denoted by $A \cap B$, is defined as a set consisting of all those points which belong to both **A and B**.

   i.e.$A \cap B = \{x: x \in A \text{ and } x \in B\}$



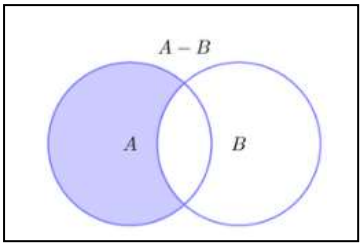
   For example: If $A = \{1, 2, 3, 5, 8\}$and $B = \{3, 5, 11\}$, then $A \cap B = \{3, 5\}$

3. **Disjoint Events:** If two events A and B have no common point, i.e. $A \cap B = \phi$, where ϕ is the null set. Then A and B are said to be disjoint events or mutually exclusive events.



4. **Relative difference of Events:** The relative difference of a event A from another event B, denoted by $A - B$ is defined as a set consisting of those elements of **A which do not belong to B.**

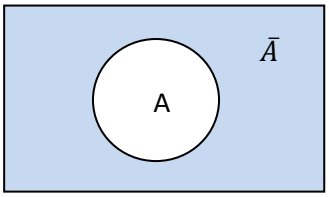
   i.e. $A - B = \{x: x \in A \text{ and } x \notin B\}$



   For example: If $A = \{1, 2, 3, 5, 8\}$and $B = \{3, 5, 11\}$, then $A - B = \{1, 2, 8\}$

5. **Complementary Event:** The complement or negative of any event A, denoted by $\bar{A}$ or A^c is a set containing all elements of the universal set S (say) other than A.

   i.e. $\bar{A} = S - A$



Classical Definition of Probability

If there are n exhaustive, mutually exclusive and equally likely events in any trial and m of them are favourable to happening of an event E, then the probability of happening of even E, denoted by $'p'$ or $P(E)$ is given by

$$p \approx P(E) = \frac{\text{Favourable number of cases}}{\text{Exaustive number of cases}}$$

or $$p = \frac{m}{n}$$

If q is the probability of non-happening of an even E, then

$$q = \frac{n-m}{n} = 1 - \frac{m}{n}$$

or $$q = 1 - p$$

Note: (i) $0 \le p \le 1, 0 \le q \le 1$

(ii) If p or $P(E) = 1$, then E is called certain event.

If p or $P(E) = 0$, then E is called impossible event.

(iii) Total probability (probability of happening and non happening) is always 1

i.e. $p + q = 1$

Examples 1.6: Two unbiased coins are tossed. What is the probability of getting at least one head?

Solution: Here possible cases are HH, HT, TH, TT and all these cases are mutually exclusive and equally likely. The cases favourable to get one head are HH, HT, TH.

Let A be the event that at least one head occurs. Then

$$P(A) = \frac{3}{4}$$

Example 1.7: Find the probability of getting both red balls, when from a bag containing 5 red and 4 black balls, two balls are drawn.

(i) with replacement
(ii) without replacement

Solution: (i) Total number of balls in the bag in both the draws = 5+4 = 9

Hence, exhaustive number of cases $= 9 \times 9 = 81$

Since there are 5 red balls in the bag.

Therefore, favourable number of cases $= 5 \times 5 = 25$

Hence, probability of both red balls $(p) = \frac{25}{81}$

(ii) Total number of ways in which 2 balls out of 9 balls can be chosen (exhaustive number of cases) $= {}^{9}C_2$

Two red balls out of 5 red balls can taken (favourable number of cases) $= {}^{5}C_2$

Hence, P (both red balls) $= \dfrac{{}^{5}C_2}{{}^{9}C_2} = \dfrac{5}{18} = 0.625$

Example 1.8: A drum contains 3 black balls, 5 red balls and 6 green balls. If 4 balls are selected at random what is the probability that the 4 selected contain

(i) No red ball?

(ii) Exactly 1 black ball?

(iii)Exactly 1 red ball and exactly 2 green balls?

Solution: Total number of balls = 14

Number of ways 4 balls can be selected = ${}^{14}C_4$

So exhaustive number of cases = ${}^{14}C_4$

(i) P (No red balls) = ${}^{9}C_4/{}^{14}C_4$

(ii) P(one black) =$({}^{3}C_1 \times {}^{11}C_3)/{}^{14}C_4$

(iii) P(1R,2G) = $({}^{5}C_1 \times {}^{6}C_2 \times {}^{3}C_1)/{}^{14}C_4$

Example 1.9: What is the probability of getting 9 cards of the same suit in one hand at a game of bridge?

Solution: One hand in the game of bridge contains 13 cards.

Therefore,

Exhaustive number of cases = ${}^{52}C_{13}$

Number of ways of getting 9 cards of one suit are = ${}^{13}C_9$

Number of ways of getting 4 cards of another suit are = ${}^{39}C_4$

There are 4 suits in a pack of cards.

Hence favourable number of cases = $4 \times {}^{39}C_4 \times {}^{13}C_9$

Therefore, probability of getting 9 cards of same suit is

$$p = \frac{4 \times {}^{39}C_4 \times {}^{13}C_9}{{}^{52}C_{13}}$$

Example 1.10: Twelve balls are distributed at random among three boxes. What is the probability that the first box will contain three balls?

Solution: Each ball can be placed in three boxes in 3 ways.

So, twelve balls can be placed in three boxes in 3^{12} ways.

Thus, exhaustive number of cases $= 3^{12}$.

Number of ways in which three balls out of twelve can be placed in I box is ${}^{12}C_3$.

Now 9 balls can be placed in 2 boxes in 2^9 ways.

Hence favourable number of cases $={}^{12}C_3 \times 2^9$

Therefore, required probability is

$$p = \frac{{}^{12}C_3 \times 2^9}{3^{12}}$$

Example 1.11: From 25 tickets, marked with the first 25 numerals, one is drawn at random. Find the chance that

(i) it is multiple of 5 or 7.

(ii) it is multiple of 3 or 7.

Solution: (i) Multiple of 5 are 5, 10, 15, 20, 25

Multiple of 7 are 7, 14, 21

Favourable number of cases = 5+3 = 8

Thus $p = \frac{8}{25}$.

(ii) Multiple of 3 are 3, 6, 9, 12, 15, 18, **21**, 24

Multiple of 7 are 7, 14, **21**

Favourable number of cases = 8+3-1=10 (1 is subtracted as 21 is common)

Thus $p = \frac{10}{25}$.

Unsolved Problems

Problem 1.1: Two unbiased dice are thrown. What is the probability that the sum of top faces is 9?

Problem 1.2: Find the probability that among two digitized numbers formed by 1, 2, 3, 4, 5, there is no repetition.

Problem 1.3: Two unbiased dice are thrown. What is the probability that both dice have the same face?

Problem 1.4: A club has four members A, B, C and D. It is required to select a Chairman and a Secretary. Assuming that one member cannot occupy both positions and selections are equally likely. What is the probability that the member A is an office bearer?

Problem 1.5: Four persons are chosen at random from a group of 3 men, 2 women and 4 children. Show that the chance of exactly two of them will be children is 10/21.

Statistical or Empirical Probability

- If the various outcomes of the trial are not equally likely.
- If the exhaustive number of cases in a trial is infinite.

Then we apply statistical or empirical probability.

Statistical Definition of Probability: If a trial is repeated a number of times under essentially homogeneous and identical conditions, then the limiting value of the ratio of the number of times the event happens to the number of trials, as the number of trials become indefinitely large, is called the probability of happening of the event.

If in n trials an event E happens m times, then probability p is

$$p = \lim_{n\to\infty} \frac{m}{n}$$

Axiomatic Definition of Probability

Let E be an experiment and S be the sample space associated with E. Then with each event $A \in S$, we associate a real number $P(A)$, called the probability of happening of any event A, and defined as

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

where $n(A)$ is the number of outcomes (elements) in A (favourable number of cases)
and $n(S)$ is the number of outcomes (elements) in S (exhaustive number of cases)

and satisfies the following properties

(i) $0 \le P(A) \le 1$

(ii) $P(S) = 1$

(iii) If A and B are mutually exclusive events, then

$$P(A \cup B) = P(A) + P(B)$$

If $A_1, A_2, ..., A_n, ...$ are pairwise mutually exclusive events, then

$$P(A_1 \cup A_2 \cup ... \cup A_{n-1} \cup A_n \cup ...) = P(A_1) + P(A_2) + ... + P(A_n) + ...$$

and for any finite n

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

Example 1.12: If an unbiased die is thrown, then find the probability of following events.

- getting an even number
- getting a number more than 4

Solution: Here sample space is given as

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

Let E_1= getting an even number ={2,4,6}$\subseteq S$

and E_2 = getting a number more than 4={5,6}$\subseteq S$

Then

$$P(E_1) = \frac{3}{6}$$

$$P(E_2) = \frac{2}{6}$$

Example 1.13: In the above **example 1.4**, find the probability of getting

- at least 2 defectives
- at most one defective
- 3 defectives

Solution:

The sample space is given as

$$S = \{DDD, DDN, DND, DNN, NDD, NDN, NND, NNN\}$$

Let A, B and C denote the above events respectively, that is

A={at least 2 defectives}= {DDD,DDN,DND,NDD}$\subseteq S$

B={at most one defective}= {DNN,NDN,NND,NNN}$\subseteq S$

C={3 defectives}={DDD}$\subseteq S$

Therefore

$$P(A) = \frac{1}{2}, P(B) = \frac{1}{2} \text{and } P(C) = \frac{1}{8}$$

Unsolved Problems

Problem 1.5: A die is rolled. What is the probability of getting a prime number?

Problem 1.6: A die is rolled once. What is the probability that (i) the number '7'coming up, (ii) number less than equal to 7.

Problem 1.7: Two coins are tossed simultaneously. Find the probability of getting (i) atleast one head, (ii) at most 1 head, (iii) no head, (iv) exactly two heads.

Problem 1.8: What is the chance that a leap year will contain 53 Sundays?

Department of Statistics & O.R.

Aligarh Muslim University Aligarh

Binomial Distribution

Bernoulli Trail and Bernoulli Distribution

Bernoulli trial is an experiment with only two possible outcomes and labeled as "Success" and "Failure".

Examples:

1. Tossing a coin (Head = Success, Tail = Failure)
2. Inspecting an item (Success=Defective, failure=Non-defective)

Let X be a discrete random variable, and

$X = 1$, if there is success in the trail

$= 0$, if failure

Further, if

$P(X=1) = p$, then $P(X=0) = 1-p$, $0 \le p \le 1$

This X is called Bernoulli random variable and its probability distribution is written as

X	:	0	1
$P(X=x)$:		$1-p$	p

which can also be written as

$$P(X=x) = p^x (1-p)^{1-x}, x = 0,1$$

The above probability distribution is called Bernoulli distribution.

Example 4.1: A die is rolled and showing up of a 1 or 6 is called a success, failure otherwise. Then

$$P(X=1) = \frac{2}{6} \text{ and } P(X=0) = \frac{4}{6}$$

Mean and Variance

$$E(X) = \sum_{x=0}^{1} x P(X=x)$$

$$= 0 \times (1-p) + 1 \times p = p$$

$$E(X^2) = \sum_{x=0}^{1} x^2 P(X=x)$$

$$= 0^2 \times (1-p) + 1^2 \times p = p$$

$$V(X) = E(X^2) - [E(X)]^2$$

$$= p - p^2 = p(1-p)$$

$$= pq,\ q = 1-p$$

Binomial Distribution

If we repeat Bernoulli trial n times and let X be the number of successes in n trials and each trial has constant probability of success (p). Then the probability distribution of number of successes (X) is called Binomial distribution.

In the definition above, we notice that the following conditions need to be satisfied for a binomial experiment:

1. There are a fixed number of n trials carried out.
2. The outcome of a given trial is either a "**success**" or "**failure**".
3. The probability of success (p) remains constant from trial to trial.
4. The trials are independent, and the outcome of a trial is not affected by the outcome of any other trial.

No. of Bernoulli Trials (n)	Outcomes	X	Probability Distribution	$P(X=x)$
1	F S	0 1	$1-p=q$ p	$p^x q^{1-x}, x=0,1$
2	FF $\left.\begin{matrix}FS\\SF\end{matrix}\right\}$ SS	0 1 2	$qq=q^2$ $\left.\begin{matrix}qp\\pq\end{matrix}\right\}=2pq$ $pp=p^2$	$\binom{2}{x}p^x q^{2-x}, x=0,1,2$
3	FFF $\left.\begin{matrix}FFS\\FSF\\SFF\end{matrix}\right\}$ $\left.\begin{matrix}SSF\\SFS\\FSS\end{matrix}\right\}$ SSS	0 1 2 3	$qqq=q^3$ $\left.\begin{matrix}qqp\\qpq\\pqq\end{matrix}\right\}=3pq^2$ $\left.\begin{matrix}ppq\\pqp\\qpp\end{matrix}\right\}=3p^2q$ $ppp=p^3$	$\binom{3}{x}p^x q^{3-x}, x=0,1,2,3$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
n	$FFSSFSF\ldots FS$ $x \to S$ $(n-x) \to F$	0 1 $\vdots$ n	x successes and $(n-x)$ failure can be arranged in $\binom{n}{x}$ ways	$\binom{n}{x}p^x q^{n-x},\ x=0,1,\ldots,n$

Therefore, finally we can define

A random variable X is said to have binomial distribution with parameters n and p if its probability mass function or probability distribution is given by

$$P(X=x)=b(x;n,p)=\begin{cases}\binom{n}{x}p^x\,q^{n-x}\ ; & x=0,1,2,\ldots,n;\ q=1-p\\ 0\ ;\ \text{otherwise}\end{cases}$$

Example: Suppose that 25% of the products of a manufacturing process are defective. Three items are selected at random, inspected, and classified as defective (D) or non-defective (N). Find the probability distribution of the number of defective items.

Solution: Here

Number of trials $(n)=3$

Probability of defective items $(p)=0.25$

Let X be the number of defective in 3 randomly selected items. Thus

$$X\sim b(3;.25)$$

Therefore, probability distribution of X is

$$P(X=x)=\binom{3}{x}(0.25)^x(0.75)^{3-x},x=0,1,2,3$$

Example: In example 4.2, find

(i) The probability of getting at least two defective items.
(ii) The probability of getting at most two defective items.

Solution: Probability distribution of number of defective is given as

$$P(X=x)=\binom{3}{x}(0.25)^x(0.75)^{3-x},x=0,1,2,3$$

X	0	1	2	3
$P(X=x)$	27/64	27/64	9/64	1/64

(i) Probability of getting atleast two defective

$$P(X\geq 2)=P(X=2)+P(X=3)=\frac{9}{64}+\frac{1}{64}=\frac{10}{64}$$

(ii) Probability of getting atmost two defective

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

$$=\frac{27}{64}+\frac{27}{64}+\frac{9}{64}=\frac{63}{64}$$

Alternatively, this can also be solved as

$$P(X\leq 2)=1-P(X>2)=1-\frac{1}{64}=\frac{63}{64}$$

Mean and Variance

We know that

$$\mu'_{(1)}(X) = E(X) = \sum_{x=0}^{n} x P(X = x)$$

$$= \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x},\ q = 1 - p$$

$$= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x q^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$= np \sum_{x'=0}^{n'} \frac{n'!}{x'!(n'-x')!} p^{x'} q^{n'-x'};\ x' = x-1, n' = n-1$$

$$= np(q+p)^{n'}$$

$$= np \quad [\because q + p = 1]$$

Now $\mu'_{(2)}(X) = E[X(X-1)] = \sum_{x=0}^{n} x(x-1) P(X = x)$

$$= \sum_{x=0}^{n} x(x-1) \binom{n}{x} p^x q^{n-x},\ q = 1 - p$$

$$= \sum_{x=0}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!} p^x q^{n-x}$$

$$= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!} p^{x-2} q^{n-x}$$

$$= n(n-1)p^2 \sum_{x'=0}^{n'} \frac{n'!}{x'!(n'-x')!} p^{x'} q^{n'-x'};\ x' = x-2, n' = n-2$$

$$= n(n-1)p^2 (q+p)^{n'}$$

$$= n(n-1)p^2$$

Therefore,

Mean $= \mu'_{(1)}(X) = E(X) = np$

Variance $= E(X^2) - [E(X)]^2 = \mu'_{(2)}(X) + \mu'_{(1)}(X) - \mu'^2_{(1)}(X)$

$$= n(n-1)p^2 + np - n^2 p^2$$

$$= np(1-p)$$

$$= npq$$

Mode

Let X be the number of success in n trials, then mode is the value of X at which probability is maximum. That is

$$P(X = x+1) \le P(X = x) \ge P(X = x+1)$$

$$\Rightarrow \quad \binom{n}{x-1} p^{x-1} q^{n-x+1} \le \binom{n}{x} p^{x} q^{n-x} \ge \binom{n}{x+1} p^{x+1} q^{n-x-1}$$

or $\quad (n+1)p - 1 \le x \le np + p$

Thus, mode is bimodal.

Example 4.4: In example 4.2, find the expected number of defective items and variance.

Solution: We know that, if $X \sim b(n; p)$, then

$$E(X) = np \text{ and } V(X) = npq$$

In our case $X \sim b(3; 0.25)$,

Therefore, expected number of defective items = $E(X) = .75$

and variance = $V(X) = 0.5625$

Example 4.5: The mean and variance of binomial distribution are 4 and 4/3 respectively. Find $P(X \ge 1)$.

Solution: Let a *r.v.* $X \sim b(n, p)$, then

$$E(X) = np \text{ and } V(X) = npq$$

We have given that

$$np = 4 \text{ and } npq = 4/3$$

$$\Rightarrow \quad q = \frac{1}{3} \text{ and } p = \frac{2}{3}$$

and $\quad n = 4 \times \frac{3}{2} = 6$

Therefore,

$$P(X = x) = \binom{6}{x} \left(\frac{2}{3}\right)^{x} \left(\frac{1}{3}\right)^{6-x}; \; x = 0, 1 \ldots, 6,$$

which gives

$$P(X \ge 1) = 1 - P(X < 1)$$

$$= 1 - P(X = 0)$$

$$=1-q^6$$

$$=1-\left(\frac{1}{3}\right)^6=0.99863$$

Example 4.6: Probability that a missile will hit the target is 0.5. How many missiles must be launched so that the probability of atleast once hitting the target exceeds 0.9?

Solution: Let n be the number of missiles launched and X is the number of missiles hitting the target. Then

$$X \sim b(n, 0.5)$$

and $$P(X=x)=\binom{n}{x}(0.5)^x(0.5)^{n-x}=\binom{n}{x}(0.5)^n,\ x=0,1,\ldots,n$$

We have given that

$$P(X \geq 1) > 0.9$$

$\Rightarrow$ $$1-P(X<1)>0.9$$

or $$1-P(X=0)>0.9$$

or $$1-(0.5)^n>0.9$$

or $$(0.5)^n<0.1 \qquad (*)$$

n	1	2	3	4
$(0.5)^n$	0.5	0.25	0.125	$0.0625<0.1$

In above table at $n=4$, (*) is satisfied.

Hence, 4 missiles to be launched.

Example 4.7: An urn contains ***a*** white balls and ***b*** blank balls. From this ***n*** balls are sampled with replacement. Find the probability that the sample contains ***x*** white balls.

Solution: Since sampling is with replacement, therefore, probability of selection of a white ball at any draw is $\frac{a}{a+b}$.

Therefore,

$$X \sim B\left(n, \frac{a}{a+b}\right)$$

and $$P(X=x)=\binom{n}{x}\left(\frac{a}{a+b}\right)^x\left(\frac{b}{a+b}\right)^{n-x},\ x=0,1,\ldots,n$$

Example 4.8: Let $X \sim B(n,p)$, then show that

$$Cov\left(\frac{X}{n}, \frac{n-X}{n}\right)=-\frac{pq}{n},\ q=1-p$$

Solution: We know that, if $X \sim b(n,p)$, then

$$P(X=x)=\binom{n}{x}p^x q^{n-x},\ x=0,1,\ldots,n$$

Thus, $E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)=\frac{1}{n}\times np=p$

$$E\left(\frac{n-X}{n}\right)=1-\frac{1}{n}E(X)=1-p=q$$

$$V\left(\frac{X}{n}\right)=\frac{1}{n^2}V(X)=\frac{1}{n^2}\times npq=\frac{pq}{n}$$

We know that

$$Cov(X,Y)=E\left[\{X-E(X)\}\{Y-E(Y)\}\right]$$

Then $Cov\left(\frac{X}{n},\frac{n-X}{n}\right)=E\left[\left\{\frac{X}{n}-E\left(\frac{X}{n}\right)\right\}\left\{\frac{n-X}{n}-E\left(\frac{n-X}{n}\right)\right\}\right]$

$$=E\left[\left\{\frac{X}{n}-p\right\}\left\{1-\frac{X}{n}-q\right\}\right]$$

$$=-E\left(\frac{X}{n}-p\right)^2$$

$$=-V\left(\frac{X}{n}\right)$$

$$=-\frac{pq}{n}$$

Example 4.9: Let $X_i, i=1,2,\ldots,k$ be independent binomial variables with parameters (n_i,p). Then show that $S_k=\sum_{i=1}^{k}X_i$ has a binomial distribution with parameters (n,p), where $n=\sum_{i=1}^{k}n_i$.

Solution: The *mgf* of S_k is

$$M_{S_k}(t)=\prod_{i=1}^{k}M_{X_i}(t)$$

$$=\prod_{i=1}^{k}(q+pe^t)^{n_i}$$

$$=(q+pe^t)^{n_1+n_2+\ldots+n_k}$$

$$=(q+pe^t)^n,$$

which is is *mgf* of binomial distribution with parameters *n* and *p*.

Now using the uniqueness property of *mgf*, $X\sim b(n,p)$.

Unsolved Problems:

Problem 4.1: The mean and variance of binomial variable X with parameters *n* and *p* are 16 and 8. Find $P(X=0)$, $P(X=1)$ and $P(X\geq 2)$.

Problem 4.2: In rolling of an unbiased die 5 or 6 is considered as success. If die is tossed 8 times. Find mean number of successes and the standard deviation.

Problem 4.3: In the long run 3 vessels out of every 100 are sunk. If 100 vessels are out, what is the probability that

(i) exactly 6 will arrive safely.
(ii) atleast 6

Problem 4.4: 10% of a population can speak English. Find the probability that in a set of 3 randomly selected persons

(i) None (ii) Atleast one (iii) Atmost one

can speak English.

Problem 4.5: A coin is tossed 10 times. How many heads are most probable to occur.

Problem 4.6: Two dice are thrown n times. Let X denotes the number of throws in which the number on the first dice exceeds the number on the second dice. What is the probability distribution of X ?

Department of Statistics & O.R.

Aligarh Muslim University, Aligarh

BA/BSc II Semester

Poisson Distribution

The Poisson distribution is used to model the number of events occurring within a given time interval.

This distribution was discovered by French mathematician and physicist Simon Denis Poisson (1781 – 1840), who published it in 1837.

Situations where Poisson distribution may occur

Poisson distribution occurs when there are events which do not occur as outcomes of a definite number of trials (unlike that in binomial) of an experiment but occur at random points of time and space, wherein our interest lies only in the number of occurrences of the event, not its non-occurrences.

Here are some situations where we can use Poisson distribution

1. Number of WhatsApp messages received on a cell phone in some unit of time.
2. Number of calls received on your mobile in some unit of time
3. Number of aircraft accidents in some unit of time
4. Number of items failed in manufacturing process during given time period
5. Number of defective items in a box of 100 items.

Definition

A discrete random variable X with countable infinite set of possible values is said to have Poisson distribution with parameter $(\lambda > 0)$, if its probability distribution is given by

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

and it is denoted by $X \sim Pois(\lambda)$.

Therefore,

$$\text{Mean } = \mu'_{(1)}(X) = E(X) = \lambda$$

$$\text{Variance } = \mu'_{(2)}(X) + \mu'_{(1)}(X) - \mu'^{2}_{(1)}(X)$$

$$= \lambda^2 + \lambda - \lambda^2 = \lambda$$

Remark 4.2: Mean and variance of Poisson distribution is same and based on its parameter.

So, for the unit time interval mean and variance is given as

$$E(X) = \lambda \text{ and } V(X) = \lambda$$

and for the time interval $(0,t)$, it becomes

$$E(X) = \lambda t \text{ and } V(X) = \lambda t$$

Remark 4.3: Mode of the Poisson distribution

- If λ is not an integer, the mode is the integral value between $(\lambda - 1)$ and λ.

- If λ is an integer, the probability at λ and $(\lambda-1)$ is same. The distribution in this case is bimodal.

Remark 4.4: Let $X \sim b(n,p)$, then

$$P(X=x)=\binom{n}{x}p^x q^{n-x}, x=0,1,2\ldots,n$$

Let $n \to \infty$ and $p \to 0$, then under these conditions, we have

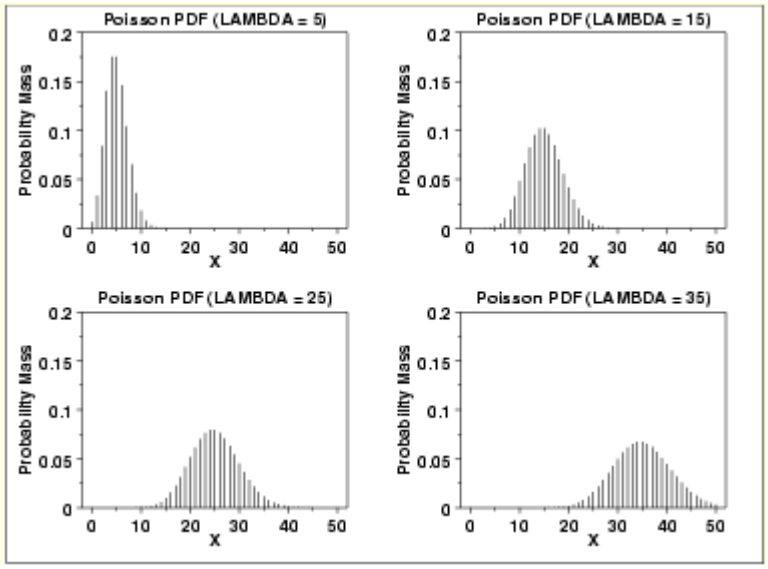
$$\lim_{n\to\infty} P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}$$

where $\lambda = np$ (Constant).

Remark 4.5: Recurrence relations for calculating the successive probabilities is given as

$$P(X=x+1)=\frac{\lambda}{(x+1)}P(X=x)$$

Remark 4.6: Shape of the probability curve is positively skewed but it tend to symmetrical for larger value of λ.



Moment Generating Function

We know that

$$M_X(t)=E[e^{tX}]=\sum_{x=0}^{\infty} e^{tx}P(X=x)$$

$$=\sum_{x=0}^{\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!}$$

$$=\sum_{x=0}^{\infty} e^{-\lambda}\frac{(\lambda e^t)^x}{x!}$$

$$= e^{-\lambda} e^{\lambda e^t}$$

$\Rightarrow \quad M_X(t) = e^{\lambda(e^t - 1)}$

Now,

$$M'_X(t) = e^{\lambda(e^t-1)} \times \lambda e^t = \lambda e^{\lambda(e^t-1)+t}$$

and $\quad M''_X(t) = \lambda e^{\lambda(e^t-1)+t}(\lambda e^t + 1)$

Thus,

$$\mu'_1 = E(X) = M'_X(0) = \lambda \qquad : \text{ Mean}$$

and $\quad \mu'_2 = E(X^2) = M''_X(0) = \lambda^2 + \lambda$

$\Rightarrow \quad \mu_2 = \mu'_2 - \mu_1'^2 = \lambda \qquad : \text{ Variance}$

Example 4.14: If X be the number of calls received at your cell phone with mean number of calls 2. Find the probability that on a randomly selected occasion there will be

i) None
ii) One
iii) Atleast one
iv) Atmost one

Calls received.

Solution: Here $X \sim Pois(2)$, therefore

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

$$P(X = 0) = \frac{e^{-2} 2^0}{0!} = e^{-2} = 0.1353$$

$$P(X = 1) = \frac{e^{-2} 2^1}{1!} = 2e^{-2} = 0.2706$$

$$P(X \geq 1) = 1 - P(X < 1)$$

$$= 1 - P(X = 0)$$

$$= 1 - e^{-2} = 0.8647$$

$$P(X \leq 1) = P(X = 0) + P(X = 1)$$

$$= e^{-2} + 2e^{-2} = 3e^{-2} = 0.4059$$

Example 4.15: Let X equal the number of typos on a printed page with a mean of 3 typos per page. What is the probability that a randomly selected page has (i) at least one typo on it, (ii) at most one typo on it?

Solution: Let X be the number of typos on a printed page, then

$$X \sim Pois(3)$$

and $$P(X=x)=\frac{e^{-3}3^x}{x!}, x=0,1,2,\ldots$$

Now,

$$P(X \geq 1)=1-P(X<1)$$

$$=1-P(X=0)$$

$$=1-e^{-3}=0.9502$$

That is, there is just over a 95% chance of finding at least one typo on a randomly selected page when the average number of typos per page is 3.

$$P(X \leq 1)=P(X=0)+P(X=1)$$

$$=e^{-3}+3e^{-3}=4e^{-3}=0.1992$$

Example 4.16: For a Poisson random variate X

$$P(X=2)=1.25P(X=1)$$

Find the probability that 'X' exceeds its mean.

Solution: Let λ is the mean, then

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

Now we have

$$P(X=2)=1.25P(X=1)$$

Implies

$$\frac{e^{-\lambda}\lambda^2}{2!}=1.25\frac{e^{-\lambda}\lambda}{1!}$$

$$\Rightarrow \lambda=2\times 1.25=2.5$$

Thus, $P(X>2.5)=1-P(X \leq 2.5)$

$$=1-\left[P(X=0)+P(X=1)+P(X=2)\right],$$

where

$$P(X=0)=\frac{e^{-2.5}2.5^0}{0!}=e^{-2.5}=0.0820$$

$$P(X=1)=2.5e^{-2.5}=0.205$$

$$P(X=2)=\frac{(2.5)^2}{2!}e^{-2.5}=0.2563$$

Example 4.17: A fisherman arrives at his favourite fishing spot. From experience he knows that the number of fish he catches per hour follows a Poisson distribution at 0.5 fish/hour. Compute the probability that he catches at least 3 fish in four hours.

Solution: Let X be the number of fish he catches.

Here value of $\lambda = 0.5$

Since time interval is $(0,4)$, therefore

$$X \sim Pois(4 \times 0.5)$$

Thus, the probability distribution of X is

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

Therefore,

$$P(X \geq 3) = 1 - P(X < 3)$$

$$= 1 - \left[P(X = 0) + P(X = 1) + P(X = 2)\right]$$

$$= 1 - [e^{-2} + 2e^{-2} + \frac{(2)^2}{2!} e^{-2}$$

$$= 1 - 5e^{-2} = 0.3235$$

Example 4.18: A manufacturer of blades knows that 5% of his product is defective. If he sales blades in packet of 100 and guarantees that not more than 10 blades will be defective. What is approximate probability that a packet will fail to meet the guaranteed quality.

Solution: We have

$$n = 100, \; p = 5\% = 0.05$$

Let X be the number of defective blades in a packet of 100 and probability of a defective blades in the packet is 0.05.

Therefore,

$$X \sim b(100, 0.05)$$

Here, since n is large and p is small, therefore binomial distribution is approximated to Poisson distribution with mean number of defective blades $\lambda = np = 100 \times 0.05 = 5$

Thus, $X \sim Pois(5)$

and $P(X = x) = \dfrac{e^{-\lambda} \lambda^x}{x!}, \; x = 0,1,2...$

Probability that a box will fail to meet the guaranteed quality is

$$P(X > 10) = 1 - P(X \leq 10) = 1 - e^{-5} \sum_{x=0}^{10} \frac{5^x}{x!}.$$

Example 4.18(a): If a discrete random variable X follows Poisson distribution with parameter value $1/2$. Find expected value of $(X+1)!$.

Solution: We have given that

$$X \sim Pois\left(\frac{1}{2}\right)$$

Then,

$$P(X=x)=\frac{e^{-\frac{1}{2}}\left(\frac{1}{2}\right)^{x}}{x!}, x=0,1,2,...$$

Therefore,

$$E[(X+1)!]=\sum_{x=0}^{\infty}(x+1)!\frac{e^{-\frac{1}{2}}\left(\frac{1}{2}\right)^{x}}{x!}$$

$$=e^{-1/2}\sum_{x=0}^{\infty}\frac{(x+1)}{2^{x}}$$

$$=e^{-1/2}\left[\sum_{x=0}^{\infty}x.\frac{1}{2^{x}}+\sum_{x=0}^{\infty}\frac{1}{2^{x}}\right]$$

$$=e^{-1/2}\left[E(X)+1+\sum_{x=1}^{\infty}\frac{1}{2^{x}}\right] \qquad (*)$$

We know that if X is the number of trials to get first success and probability of getting success at each trial is $1/2$, then $X \sim Geo(1/2)$

$$P(X=x)=\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^{x-1}=\frac{1}{2^{x}}, x=1,2,...$$

Here $\sum_{x=1}^{\infty}\frac{1}{2^{x}}=1$ and $E(X)=\frac{1}{(1/2)}=2$

Therefore from (*)

$$E[(X+1)!]=e^{-1/2}[2+1+1]=4e^{-1/2}=0.7356$$

Unsolved Problems:

Problem 4.7: If X has a Poisson distribution with parameter λ, and if $P(X=0)=0.2$. Find $P(X>2)$.

Problem 4.8: The number of industrial injuries per working week in a factory is known to follow a Poisson distribution with mean 0.5. Find the probability that

a) In a given week there will be

   i) less than 2 accidents
   ii) more than 2 accidents

b) in a three-weekperiod, there will be no accidents.

Problem 4.9: Let X and Y are independent Poisson variables, such that

$$P(X=1)=P(X=2)$$
$$P(Y=2)=P(Y=3)$$

Find $Var(X-2Y)$.

Department of Statistics & O.R.

Aligarh Muslim University, Aligarh

Normal Distribution

A continuous random variable X is said to have normal distribution with parameters μ and σ^2, when its probability density function (*pdf*) is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty;\ -\infty < \mu < \infty,\ \sigma > 0 \qquad (4.24)$$

In terms of notation, we generally write like

$$X \sim N\left(\mu, \sigma^2\right)$$

- That means X follows normal distribution with parameters μ and σ^2.
- μ and σ are called location and scale parameters, respectively.

Special Cases:

1. If $\mu = 0$, then

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}, -\infty < x < \infty$$

is known as Gaussian distribution (or error distribution)

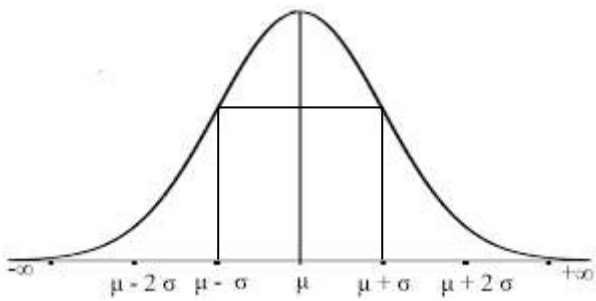
2. If $\mu = 0, \sigma^2 = 1$, then

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < \infty$$

is known as standard normal distribution.

Properties of the Normal Distribution

1) The curve of the normal distribution is bell shaped and it is symmetrical about mean μ.



2) The curve decreases as we move away from μ. There is a point of inflexion at $X = \mu \pm \sigma$.

   Point of inflexion: The point at which the concavity changes is called the point of inflexion.

3) The location of the normal distribution depends on μ and its shape depends on σ.

5) Mean and Variance

We have

$$\mu_1' = E(X) = \int_{-\infty}^{\infty} x\, f(x)dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Let $\frac{x-\mu}{\sigma} = z \Rightarrow dx = \sigma dz$, then

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z) e^{-\frac{1}{2}z^2} dz$$

$$= \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz + \frac{1}{\sqrt{2\pi}} \sigma \int_{-\infty}^{\infty} z\, e^{-z^2/2} dz$$

[Since $\int_{-\infty}^{\infty} z\, e^{-z^2/2} dz = \int_{-\infty}^{\infty} g(z)dz = 0$,

as $g(z)$ has property $g(z) = -g(-z)$, and hence g is odd function.

and $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz = 1$]

Therefore,

$$\text{Mean} = E(X) = \mu$$

Again

$$\mu_2' = E(X^2) = \int_{-\infty}^{\infty} x^2\, f(x)dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Let $\frac{x-\mu}{\sigma} = z \Rightarrow dx = \sigma dz$, then

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z)^2 e^{-\frac{1}{2}z^2} dz$$

$$= \mu^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz + \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz + 2\mu\sigma \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\, e^{-z^2/2} dz \qquad (4.27)$$

Now consider the second term of (4.27)

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} z^2 e^{-z^2/2} dz \quad \text{[due to symmetry]}$$

From (4.25) we have

$$\Gamma p = 2^{1-p} \int_0^{\infty} z^{2p-1} e^{-z^2/2} dz$$

Let $p = 3/2$, then

$$\Gamma(3/2) = \frac{1}{\sqrt{2}} \int_0^\infty z^2 e^{-z^2/2} dz$$

$$\frac{\sqrt{\pi}}{2} = \frac{1}{\sqrt{2}} \int_0^\infty z^2 e^{-z^2/2} dz \qquad [\text{Since } \Gamma(3/2) = \frac{1}{2}\Gamma(1/2) = \frac{\sqrt{\pi}}{2}]$$

or $$\int_0^\infty z^2 e^{-z^2/2} dz = \frac{\sqrt{2\pi}}{2}$$

Thus,

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty z^2 e^{-z^2/2} dz = \frac{2}{\sqrt{2\pi}} \times \frac{\sqrt{2\pi}}{2} = 1$$

Therefore from (4.27), we get

$$\mu_2' = E(X^2) = \mu^2 + \sigma^2$$

Now

$$\text{Variance} = V(X) = \mu_2' - \mu_1'^2$$

$$= \mu^2 + \sigma^2 - \mu^2 = \sigma^2$$

6) Moment Generating Function

We have

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^\infty e^{tx} f(x) dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^\infty e^{tx} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Let $\frac{x-\mu}{\sigma} = s \Rightarrow x = \mu + \sigma s$ and $dx = \sigma ds$

Therefore,

$$M_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{t(\mu+\sigma s)} e^{-s^2/2} ds$$

$$= \frac{e^{t\mu}}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}(s^2 - 2t\sigma s)} ds$$

$$= \frac{e^{t\mu}}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}(s^2 - 2t\sigma s + t^2\sigma^2 - t^2\sigma^2)} ds$$

$$= e^{t\mu + \frac{t^2\sigma^2}{2}} \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}(s - t\sigma)^2} ds$$

Again let $s-\sigma t = z \Rightarrow ds = dz$

Therefore,

$$M_X(t) = e^{t\mu + \frac{t^2\sigma^2}{2}} \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz$$

or $$M_X(t) = e^{t\mu + \frac{t^2\sigma^2}{2}} \qquad \left[\text{Since } \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = 1\right]$$

Further,

$$M'_X(t) = e^{t\mu + \frac{t^2\sigma^2}{2}} (\mu + \sigma^2 t)$$

and $$M''_X(t) = e^{\left(t\mu + \frac{t^2\sigma^2}{2}\right)} (\mu + \sigma^2 t)^2 + e^{\left(t\mu + \frac{t^2\sigma^2}{2}\right)} \sigma^2 \ ,$$

which gives

$$\mu'_1 = M'_X(0) = \mu$$

and $$\mu'_2 = M''_X(0) = \mu^2 + \sigma^2$$

Thus,

$$\text{Mean} = \mu'_1 = \mu$$

$$\text{Variance} = V(X) = \mu'_2 - \mu'^2_1$$

$$= \mu^2 + \sigma^2 - \mu^2 = \sigma^2$$

7) In normal distribution mean, median and mode coincide.

12) **Area properties:**

The total area under curve is unity (that is 1). Hence

(i) $P(X > \mu) = P(X < \mu) = \frac{1}{2}$

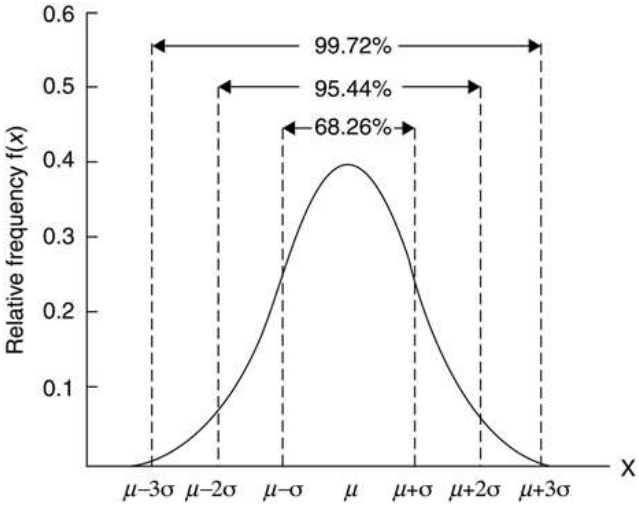
(ii) $P(\mu < X < \mu + a) = P(\mu - a < X < \mu)$

If $X \sim N(\mu, \sigma^2)$, then

$P(\mu - \sigma < X < \mu + \sigma) = .6826$

$P(\mu - 2\sigma < X < \mu + 2\sigma) = .9544$

$P(\mu - 3\sigma < X < \mu + 3\sigma) = .9972$



and subsequently

Example 4.32: If a continuous random variable X follows normal distribution with mean 12 and variance 16. Compute the following probabilities

(i) $P(X \geq 20)$, (ii) $P(X \leq 24)$, (iii) $P(X \leq 20)$, (iv) $P(0 \leq X \leq 12)$ and

(v) $P(16 \leq X \leq 24)$

Solution: We have given that

$X \sim N(12,16)$, which implies

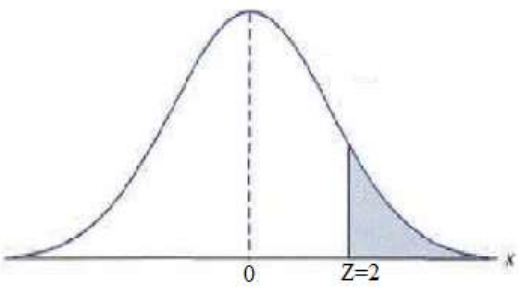
$$E(X) = \mu = 12$$

$$V(X) = \sigma^2 = 16$$

(i) $P(X \geq 20) = P\left(\frac{X-\mu}{\sigma} \geq \frac{20-\mu}{\sigma}\right)$

$$= P\left(Z \geq \frac{20-12}{4}\right)$$

$$= P(Z \geq 2)$$



$$= P(0 \leq Z < \infty) - P(0 \leq Z \leq 2)$$

$$= 0.5 - \left(\frac{.9544}{2}\right)$$

$$= 0.5 - .4772 = 0.0228$$

or

$$P(Z \geq 2) = 1 - P(-\infty \leq Z \leq 2)$$

$$= 1 - 0.9772 = 0.0228$$

(ii) $P(X \leq 24) = P\left(\frac{X-\mu}{\sigma} \leq \frac{24-\mu}{\sigma}\right)$

$$= P\left(Z \leq \frac{24-12}{4}\right)$$

$$= P(Z \leq 3)$$

$$= P(0 \leq Z < \infty) + P(0 \leq Z \leq 3)$$

$$= 0.5 + \left(\frac{.9972}{2}\right)$$

$$= 0.5 + .4986 = 0.9986$$

or $P(Z \leq 3) = P(-\infty < Z \leq 3)$

$$= .9986$$

(iii) $P(X \leq 20) = P\left(\frac{X-\mu}{\sigma} \leq \frac{24-\mu}{\sigma}\right)$

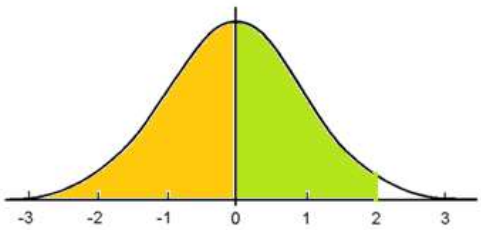
$$= P\left(Z \leq \frac{20-12}{4}\right)$$

$$= P(Z \le 2)$$

$$= P(0 \le Z < \infty) + P(0 \le Z \le 2)$$

$$= 0.5 + \left(\frac{.9544}{2}\right)$$

$$= 0.5 + .4722 = 0.9722$$



or

$$P(X \le 20) = 1 - P(X > 20)$$

$$= 1 - .0228 = 0.9772$$

(iv) $P(0 \le X \le 12)$

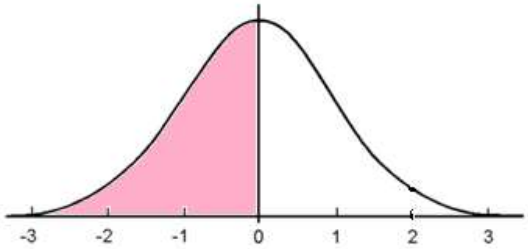
$$= P\left(\frac{0-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{12-\mu}{\sigma}\right)$$

$$= P\left(\frac{0-12}{4} \le Z \le \frac{12-12}{4}\right)$$

$$= P(-3 \le Z \le 0)$$

$$= \left(\frac{.9972}{2}\right)$$

$$= 0.4986$$



(v) $P(16 \le X \le 24)$

$$= P\left(\frac{16-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{24-\mu}{\sigma}\right)$$

$$= P\left(\frac{16-12}{4} \le Z \le \frac{24-12}{4}\right)$$

$$= P(1 \le Z \le 3)$$

$$= P(-\infty < Z \le 3) - P(-\infty < Z \le 1)$$

$$= 0.4986 - 0.3413 = 0.1573$$

Example 4.33: If $X \sim N(\mu, \sigma^2)$, then show that:

$$Y = \left(\frac{X-\mu}{\sigma}\right)^2 = Z^2$$

is distributed as chi-square *r.v.* with 1 degree of freedom.

Solution: Let $G(y)$ be the *df* of *Y*, then

$$G(y) = P(Y \le y)$$

$$= P(Z^2 \le y)$$

$$= P(-\sqrt{y} < Z < \sqrt{y})$$

$$= 2P(0 < Z < \sqrt{y}) \quad \text{[due to symmetry]}$$

$$= 2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

Let $z^2 = t \Rightarrow dz = \frac{1}{2\sqrt{t}} dt$, then

$$G(y) = \frac{2}{\sqrt{2\pi}} \int_0^y \frac{1}{2\sqrt{t}} e^{-\frac{1}{2}t} dt$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^y t^{(1/2)-1} e^{-(1/2)t} dt$$

$$\frac{\partial}{\partial y} G(y) = g(y) = \frac{1}{\sqrt{2\pi}} y^{(1/2)-1} e^{-(1/2)y}$$

$$\Rightarrow \qquad g(y) = \frac{1}{\Gamma(1/2)\, 2^{(1/2)}} y^{(1/2)-1} e^{-(1/2)y} \quad \text{[since } \sqrt{\pi} = \Gamma(1/2)\text{].}$$

This is the *pdf* of gamma distribution with $k = 1/2$ and $\lambda = 1/2$, and is called chi-square (χ^2) distribution with 1 degree of freedom. We write $Y \sim \chi_1^2$.

[Note: This problem can also be solved using *mgf* , as *mgf* uniquely determines the *df*].

Example 4.34: If $X \sim N(\mu, \sigma^2)$, then show that the even moments are

$$\mu_{2n} = 1.3.5....(2n-1)\sigma^{2n}, n = 1, 2, ...$$

Solution: By definition

$$\mu_{2n} = \int_{-\infty}^{\infty} (x-\mu)^{2n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Let $\frac{x-\mu}{\sigma} = z \Rightarrow dx = \sigma dz$, then

$$\mu_{2n} = \int_{-\infty}^{\infty} (z\sigma)^{2n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}} \times 2\int_0^\infty z^{2n} e^{-\frac{1}{2}z^2} dz$$

$$= \frac{2\sigma^{2n}}{\sqrt{2\pi}} \int_0^\infty z^{(2n+1)-1} e^{-\frac{1}{2}z^2} dz$$

$$= \frac{2\sigma^{2n}}{\sqrt{2\pi}} \times \frac{\Gamma[n+(1/2)]}{2^{1-n-(1/2)}}$$

$$= \frac{2^{n+(1/2)}\sigma^{2n}}{\sqrt{2\pi}} \times \Gamma[n+(1/2)].$$

Now

$$n = 1 \Rightarrow \mu_2 = \sigma^2$$

$$n = 2 \Rightarrow \mu_2 = 3\sigma^4$$

$$\vdots$$

$$n = n \Rightarrow \mu_{2n} = 1.3.5....(2n-1)\sigma^{2n}, n = 1, 2, ...$$

Unsolved problems:

Problem 4.15: Given a standard normal distribution, find the area under the curve that lies:

(a) to the right of z = 1.84.
(b) between z = -1.97 *and* z = 0.86.

Problem 4.16: Given a standard normal distribution, find the value of k such that

(a) $P(Z > k) = 0.3015$

(b) $P(k < Z < -0.18) = 0.4197$

Problem 4.17: Given the normally distributed variable X with mean 18 and standard deviation 2.5, find

(a) $P(X > 20)$
(b) $P(17 < X < 21)$
(c) the value of k such that $P(X > k) = 0.1814$

Problem 4.18: If X is normally distributed with mean 30 and standard deviation 5. Compute following probabilities

(i) $P(26 \le X \le 40)$, (ii) $P(|X - 30| > 5)$

Problem 4.19: A project yields an average cash flow of Rs.500 lakhs with a standard deviation of Rs. 60 lakhs. Calculate the probability of the following events:

(a) the cash flow is more than Rs. 500 lakhs.
(b) the cash flow lies between Rs. 380 lakhs and Rs. 620 lakhs.

Problem 4.20: In certain examination test 2000 students appeared in statistics. The average marks obtained were 50% and the standard deviation was 5%. How many students do you expect to obtain more than 60% marks, supposing that the marks are normally distributed.

Problem 4.21: In normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution.

Department of Statistics & O.R.

Aligarh Muslim University Aligarh

Correlation

- Simple Correlation
- Partial Correlation
- Multiple Correlation

Correlation

Correlation measures how variables are related. The correlation coefficient is used to quantify the strength and the direction of the linear relationship between two variables and usually denoted by r. The linear correlation coefficient is sometimes referred to as the ***Pearson product moment correlation coefficient*** in honour of its developer Professor Karl Pearson.

Pearson's correlation coefficient, used when

- Data is quantitative.
- Distribution of variables is normal.
- Linear relationship between two variables.
- All observations should be independent.

The mathematical formula for computing Karl Pearson coefficient of correlation between two variables X and Y is:

$$r = r_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\frac{1}{n}\sum_{i}^{n}(X_i - \bar{X})^2\right)\left(\frac{1}{n}\sum_{i}^{n}(Y_i - \bar{Y})^2\right)}}$$

$$= \frac{\sum UV}{\sqrt{\sum U^2 \sum V^2}}$$

Where $U = X_i - \bar{X}$ and $V = Y_i - \bar{Y}$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

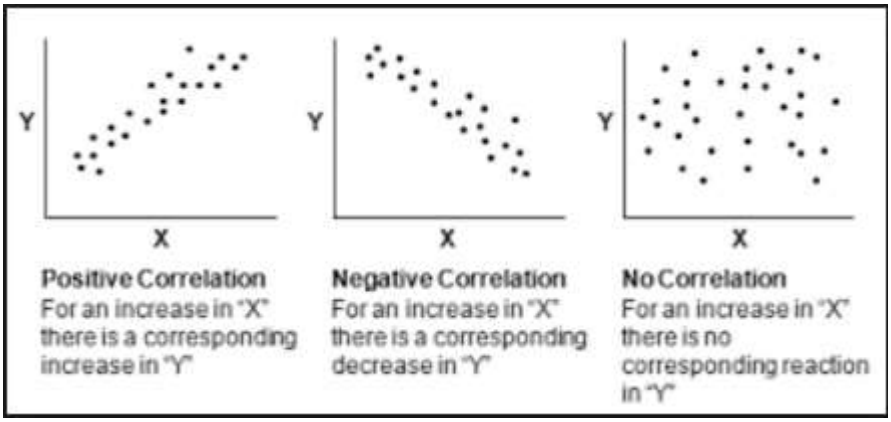
n is the number of pairs of data

Alternate formula:

$$r = \frac{\frac{1}{n}\sum XY - \bar{X}\bar{Y}}{\sqrt{\left(\frac{1}{n}\sum X^2 - \bar{X}^2\right)\left(\frac{1}{n}\sum Y^2 - \bar{Y}^2\right)}}$$

$$= \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\{n\sum X^2 - (\sum X)^2\}\{n\sum Y^2 - (\sum Y)^2\}}}$$

Scatter Diagram or Scatter Plot



Remarks:

1. The value of r is such that $-1 \le r \le +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.
2. **Positive Correlation:** If two variables X and Y deviate in same direction, that is the increase/decrease in one variable results in corresponding increase/decrease in other, the correlation is said to be positive. For example height and weight of group of persons.
3. **Negative Correlation:** If two variables X and Y deviate in opposite direction, that is the increase/decrease in one variable results in corresponding decrease/increase in other, the correlation is said to be negative. For example pressure and volume of gas.
4. **No Correlation:** The two variables X and Y is said to be uncorrelated if and increase/decrease in one results in no change in other. For example weight and *IQ*.
5. **Perfect Correlation:** The correlation is said to be perfect if the deviation in one variable followed by the corresponding proportional deviation in other. The value $r = +1$ indicates perfect positive correlation whereas $r = -1$ means perfect negative correlation.
6. The correlation coefficient r is a dimensionless quantity; that means it does not depend on the units employed.
7. Note that
   - If $r = 0$ this means no association or correlation between the two variables.
   - If $0 < r < 0.25 / -0.25 < r < 0$: weak **positive/negative** correlation.
   - If $0.25 \le r < 0.75 / -0.75 < r \le -0.25$: intermediate **positive/negative** correlation.
   - If $0.75 \le r < 1 / 1 < r \le -0.75$: strong **positive/negative** correlation.
   - If $r = +1 / r = -1$: perfect **positive/negative** correlation.

Properties:

1. Correlation coefficient lies between -1 to $+1$. That is

$$-1 \le r \le +1.$$

2. Correlation coefficient is independent of origin and scale.

3. **If two variables are independent, then correlation between then is zero but converse is not true.**

The above can also be illustrated with the help of following example, let

***X*:**	-3	-2	-1	1	2	3	$\sum \boldsymbol{X} = \mathbf{0}$
***Y*:**	9	4	1	1	4	9	$\sum Y = 28$
***XY*:**	-27	-8	-1	1	8	27	$\sum XY = 0$

Here

$$\bar{X} = \frac{1}{n}\sum_i X_i = 0$$

$$Cov(X,Y) = \frac{1}{n}\sum_i X_i Y_i - \bar{X}\bar{Y}$$
$$= 0 - 0 = 0,$$

which implies

$$r_{XY} = 0$$

But here we may observe the X and Y are related by the relation $Y = X^2$.

Example 3.1: Weight and Systolic Blood Pressure of 10 patient are given below. Find product moment correlation coefficient between Wt. and SBP.

Wt. (kg)	67	69	85	83	74	81	97	92	114	85
SBP (mmHg)	120	125	140	160	130	180	150	140	200	130

Solution:

$\boldsymbol{X}$	$\boldsymbol{Y}$	$(\boldsymbol{X}-\bar{\boldsymbol{X}})$	$(\boldsymbol{X}-\bar{\boldsymbol{X}})^2$	$(\boldsymbol{Y}-\bar{\boldsymbol{Y}})$	$(\boldsymbol{Y}-\bar{\boldsymbol{Y}})^2$	$(\boldsymbol{X}-\bar{\boldsymbol{X}})(\boldsymbol{Y}-\bar{\boldsymbol{Y}})$
67	120	-17.70	313.29	-27.5	756.25	486.75
69	125	-15.70	246.49	-22.50	506.25	353.25
85	140	0.30	0.09	-7.50	56.25	-2.25
83	160	-1.70	2.89	12.50	156.25	-21.25
74	130	-10.70	114.49	-17.50	306.25	187.25
81	180	-3.70	13.69	32.50	1056.25	-120.25
97	150	12.30	151.29	2.50	6.25	30.75
92	140	7.30	53.29	-7.50	56.25	-54.75
114	200	29.30	858.49	52.50	2756.25	1538.25
85	130	0.30	0.09	-17.50	306.25	-5.25
$\sum \boldsymbol{X} =$ **847**	$\sum \boldsymbol{Y} =$ **1475**		$\sum(\boldsymbol{X}-\bar{\boldsymbol{X}})^2 =$ $\mathbf{1754.10}$		$\sum(\boldsymbol{Y}-\bar{\boldsymbol{Y}})^2 =$ $\mathbf{5962.50}$	$\sum(\boldsymbol{X}-\bar{\boldsymbol{X}})(\boldsymbol{Y}-\bar{\boldsymbol{Y}}) =$ $\mathbf{2392.50}$

Here

$$\bar{X} = \frac{\sum X}{n} = \frac{847}{10} = 84.7 \text{ and } \bar{Y} = \frac{\sum Y}{n} = \frac{1475}{10} = 147.5$$

Therefore,

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{(\sum_i^n (X_i - \bar{X})^2)(\sum_i^n (Y_i - \bar{Y})^2)}} = \frac{2392.50}{\sqrt{1754.10 \times 5962.5}} = 0.74$$

Example 3.2: A sample of 6 children was selected, data about their age in weeks and weight in kilograms was recorded as shown in the following table . It is required to find the correlation between age and weight.

S. No.	Age (years)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

Solution:

S. No.	Age (years) X	Weight (Kg) Y	XY	X^2	Y^2
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
	$\sum X = 41$	$\sum Y = 66$	$\sum XY = 461$	$\sum X^2 = 291$	$\sum Y^2 = 742$

Therefore,

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\{n\sum X^2 - (\sum X)^2\}\{n\sum Y^2 - (\sum Y)^2\}}}$$

$$= \frac{6 \times 461 - 41 \times 66}{\sqrt{\{6 \times 291 - (41)^2\}\{6 \times 742 - (66)^2\}}} = .76$$

That means strong positive correlation between Age and Weight.

Example 3.3: In the following table age of the students and their yoga habits are given. Calculate correlation between age of the students and their yoga habits.

Age	15	16	17	18	19	20
No. of Students	250	200	150	120	100	80
Students do Yoga	200	150	90	48	30	12

Solution:

Age (X)	% of Students	$X-\bar{X}=U$	$Y-\bar{Y}=V$	U^2	V^2	UV
15	**80**	-2.5	30	6.25	900	-75
16	**75**	-1.5	25	2.25	625	-37.5
17	**60**	-0.5	10	0.25	100	-5
18	**40**	0.5	-10	0.25	100	-5
19	**30**	1.5	-20	2.25	400	-30
20	**15**	2.5	-35	6.25	1225	-87.5
$\sum X=105$	$\sum Y=300$			$\sum U^2=17.5$	$\sum V^2=3350$	$\sum UV=-240$

$$\bar{X}=\frac{1}{n}\sum X=\frac{105}{6}=17.5$$

$$\bar{Y}=\frac{1}{n}\sum Y=\frac{300}{6}=50$$

$$r=\frac{\sum UV}{\sqrt{(\sum U^2)(\sum V^2)}}=\frac{-240}{\sqrt{17.5\times 3350}}=-0.992$$

The above analysis shows that there is strong negative correlation between age of the students and their Yoga habits.

Practice Exercises:

1. Calculate Karl Pearson coefficient of correlation between X and Y for the following data:

X:	1	2	3	4	5	7	8	10
Y:	2	6	8	10	14	16	18	20

2. Anxiety Level and Test Score of 6 High School students are recorded as below:

Anxiety Level	10	8	2	1	5	6
Test Score	2	3	9	7	6	5

Find correlation between Anxiety level and Test score.

3. Consider the following data

X:	1	2	3	4	5	7	8	10
Y:	2	6	8	10	14	16	18	20

i) Calculate Karl Pearson coefficient of correlation between X and Y.

ii) If 4 is subtracted from each value of X and each value of Y is divided by 2, then what will be correlation between new series of X and Y.

4. From the following data compute the coefficient of correlation between X and Y for 15 set of observations:

	X	Y
Arithmetic Means	25	18
Sum of Square of Deviations from Means	136	138

Partial Correlation

Partial correlation analysis involves studying the linear relationship between two variables after excluding the effect of one or more independent factors. For example, suppose that there are three variables Sale (X_1), Price (X_2) and Production (X_3) and one wants to find the linear relationship between Sale (X_1) and Price (X_2) after controlling the Production (X_3).

Coefficient of Partial Correlation

Partial correlation coefficient between X_1 and X_2 after controlling the effect of X_3 and usually denoted by $r_{12.3}$ is given as

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

Similarly,

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} \text{ , as } r_{23} = r_{32}$$

and

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

Multiple Correlation

Sometimes we find interrelationship between many variables and value of one variable may be influenced by many others. For example the yield of crop per hectare, say (X_1) depends upon quality of seed (X_2), fertility of soil (X_3), fertilizer used (X_4), irrigation facilities (X_5), weather condition (X_6) and so on. When someone is interested in studying joint effect of variables upon a variable not included in that group, our study is that of multiple correlation and multiple linear regression analysis.

Coefficient of Multiple Correlations

Consider a trivariate distribution in which each of the variable X_1, X_2 and X_3 has n observations.

The multiple correlation coefficient of X_1 on X_2 and X_3, usually denoted by $R_{1.23}$ is given as

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

This is also called simple correlation between X_1 and joint effect of X_2 and X_3 on X_1.

Similarly, the multiple correlation coefficient of X_2 on X_1 and X_3 is given as

$$R^2_{2.13} = \frac{r_{12}^2 + r_{23}^2 - 2\, r_{12} r_{23} r_{13}}{1 - r_{13}^2}$$

Properties of Multiple Correlation

i) $0 \leq R_{1.23} \leq 1$

ii) If $R_{1.23} = 1$, then association is perfect, and the regression residuals are zero.

iii) If $R_{1.23} = 0$, then all total and partial correlation involving X_1 are zero. So, X_1 is totally uncorrelated with all other variables.

iv) $R_{1.23}$ is not less than any total correlation coefficient

$$R_{1.23} \geq r_{12}, r_{13}, r_{23}$$

Example 3.10: A study was conducted to know role of hours spent on revision and anxiety level on exam score. The following data were recorded

Student No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Score	62	58	52	55	75	82	38	55	48	68	62	62	72	58
Hours	40	31	35	26	51	48	25	37	30	44	32	40	61	35
Anxiety	40	65	34	91	46	52	48	61	34	74	54	61	26	13

Compute

i) The correlation between exam score and hours of revision after controlling the effect of anxiety level.

ii) The correlation between exam scores and anxiety level after controlling the effect of hours of revision.

iii) The joint effect of hours of revision and anxiety level on exam score.

Solution:

Score (X_1)	Hours (X_2)	Anxiety (X_3)	X_1^2	X_2^2	X_3^2	X_1X_2	X_2X_3	X_1X_3
62	40	40	3844	1600	1600	2480	1600	2480
58	31	65	3364	961	4225	1798	2015	3770
52	35	34	2704	1225	1156	1820	1190	1768
55	26	91	3025	676	8281	1430	2366	5005
75	51	46	5625	2601	2116	3825	2346	3450
82	48	52	6724	2304	2704	3936	2496	4264
38	25	48	1444	625	2304	950	1200	1824
55	37	61	3025	1369	3721	2035	2257	3355
48	30	34	2304	900	1156	1440	1020	1632
68	44	74	4624	1936	5476	2992	3256	5032
62	32	54	3844	1024	2916	1984	1728	3348
62	40	61	3844	1600	3721	2480	2440	3782
72	61	26	5184	3721	676	4392	1586	1872
58	35	13	3364	1225	169	2030	455	754
847	**535**	**699**	**52919**	**21767**	**40221**	**33592**	**25955**	**42336**

First we shall compute the following simple correlations

$$Corr(X_1, X_2) = r_{12} = \frac{n\sum X_1X_2 - \left(\sum X_1\right)\left(\sum X_2\right)}{\sqrt{\left\{n\sum X_1^2 - \left(\sum X_1\right)^2\right\}\left\{n\sum X_2^2 - \left(\sum X_2\right)^2\right\}}}$$

$$= \frac{14\times 33592 - 847\times 535}{\sqrt{\left\{14\times 52919 - (847)^2\right\}\left\{14\times 21767 - (535)^2\right\}}} = 0.8226$$

$$Corr(X_1, X_3) = r_{13} = \frac{n\sum X_1X_3 - \left(\sum X_1\right)\left(\sum X_3\right)}{\sqrt{\left\{n\sum X_1^2 - \left(\sum X_1\right)^2\right\}\left\{n\sum X_3^2 - \left(\sum X_3\right)^2\right\}}}$$

$$= \frac{14\times 42336 - 847\times 699}{\sqrt{\left\{14\times 52919 - (847)^2\right\}\left\{14\times 40221 - (699)^2\right\}}} = 0.0156$$

$$Corr(X_2, X_3) = r_{23} = \frac{n\sum X_2X_3 - \left(\sum X_2\right)\left(\sum X_3\right)}{\sqrt{\left\{n\sum X_2^2 - \left(\sum X_2\right)^2\right\}\left\{n\sum X_3^2 - \left(\sum X_3\right)^2\right\}}}$$

$$= \frac{14\times 25955 - 535\times 699}{\sqrt{\left\{14\times 21767 - (535)^2\right\}\left\{14\times 40221 - (699)^2\right\}}} = -0.2853$$

(i) The partial correlation between exam score and hours of revision after controlling the effect of anxiety level is

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} = \frac{0.8226 - 0.0156\times(-0.2853)}{\sqrt{(1-(0.0156)^2)(1-(-0.2853)^2)}} = 0.863$$

(ii) The partial correlation between exam score and anxiety level after controlling the effect of hours of revision

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} = \frac{0.0156 - 0.8226\times(-0.2853)}{\sqrt{(1-(0.8226)^2)(1-(-0.2853)^2)}} = 0.459$$

iii) The joint effect of hours of revision and anxiety level on exam score.

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

$$= \frac{(0.8226)^2 + (0.0156)^2 - 2\times 0.8226\times 0.0156\times(-0.2853)}{1-(-0.2853)^2}$$

$$= \frac{0.6842}{0.9186} = 0.7448$$

TEST BASED ON $t-$DISTRIBUTION

Suppose Y and Z an independent random variables, Y has a chi-square distribution with $(n-1)$ degree of freedom (df) and Z has the standard normal distribution, then the distribution of new variable $Z/\sqrt{Y/(n-1)}$ is called t with $(n-1)$ df. It has probability density function (pdf)

$$f(t)=\frac{1}{(n-1)^{1/2}B\left(\frac{1}{2},\frac{n-1}{2}\right)}\left(1+\frac{t^2}{n-1}\right)^{-n/2}, \quad -\infty<t<\infty.$$

Consider

$$t = Z\sqrt{Y/(n-1)}\text{, where} \quad Z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1) \text{ and}$$

$$Y=\sum_i\left(\frac{x_i-\bar{x}}{\sigma}\right)^2=\frac{1}{\sigma^2}\sum(x_i-\bar{x})^2=\frac{1}{\sigma^2}(n-1)s^2\sim\chi^2_{n-1}, \quad , \; s^2=\frac{1}{n-1}\sum(x_i-\bar{x})^2 .$$

Thus

$$t=\frac{\sqrt{n}\,(\bar{x}-\mu)/\sigma}{s/\sigma}=\frac{\sqrt{n}\,(\bar{x}-\mu)}{s}\sim t_{n-1}.$$

Assumptions for students $t-$test

i) The parent population from which the sample is drawn is normal.

ii) The sample observations are independent i.e., the sample is random.

iii) The population standard deviation σ is unknown.

Test for single mean

When $n<30$ and σ^2 unknown, then in this case we use $t-$distribution instead of standard normal distribution.

Suppose we want to test

$H_0:\mu=\mu_0$ (specified), we may therefore, use the statistic, under H_0

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}.$$

For $H_A:\mu<\mu_0$, H_0 is rejected if for the given sample

$$t<t_{\alpha,n-1}$$

For $H_A:\mu<\mu_0$, H_0 is rejected if for the given sample

$$t>t_{\alpha,n-1}$$

For $H_A:\mu\neq\mu_0$, H_0 is rejected if

$$|t|>t_{\alpha/2,n-1}.$$

Confidence limit for population mean μ

By definition

$$\Pr(|t| \le t_{\alpha/2,n-1}) = 1-\alpha \qquad \text{or} \quad \Pr\left(\left|\frac{\bar{x}-\mu}{s/\sqrt{n}}\right| \le t_{\alpha/2,n-1}\right) = 1-\alpha$$

$$\Rightarrow \Pr\left(\bar{x} - t_{\alpha/2,n-1}\, s/\sqrt{n} \le \mu \le \bar{x} + t_{\alpha/2,n-1}\, s/\sqrt{n}\right) = 1-\alpha .$$

$100(1-\alpha)\%$ of the cases, this interval $\bar{x} \pm t_{\alpha/2,n-1}\, s/\sqrt{n}$ include μ.

Exercise 1: 10 individuals are chosen at random from a normal population and the heights are found to be in inches 63, 63, 66, 67, 68, 69, 68, 70, 71 and 71, in the light of these data, discuss the suggestion that the mean height in the population is 66 inches.

Solution: Here $\bar{x} = \text{sample mean} = 67.6$ inches, $n = 10$, $s = 2.9135$ inches

H_0: $\mu = 66$ inches

H_A : $\mu \ne 66$ inches.

Test statistic

$$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{67.8-66}{3.011/\sqrt{10}} = 1.647 \text{ approx.}$$

Also, $n-1 = \text{number of } df = 9$, for 9 df, we get from the table,

$$t_{0.025,9} = 2.262.$$

As the calculated value of $|t| = 1.65$, is less than $t_{0.025,9} = 2.262$ (the difference between $\bar{x}$ and μ may be due to fluctuations of random sampling) and the H_0, may be retained at 5% level of significance.

Exercise 2: The weights in lb. at birth for 15 babies born in a hospital are given below,

6.2, 6.7, 7.1, 6.9, 7.5, 5.7, 4.8, 6.8, 7.6, 7.8, 8.1, 5.0, 5.8, 7.9, 8.5.

Find the limits between which the mean weight at birth for all such babies is likely to lie.

Solution: Assume that the given observations are from normal population, $100(1-\alpha)\%$ confidence limits to μ will be $\bar{x} \pm t_{\alpha/2,n-1}\, s/\sqrt{n}$.

In the present case

$$n = 15,\ \sum x_i = 102.4,\ \sum x_i^2 = 716.88$$

and hence $\bar{x} = 6.827$lb. and $s = 1.126$lb.

From table $t_{0.005,14} = 2.977$.

Hence 99% confidence limits to μ be

$6.827 \pm 2.977\dfrac{1.126}{\sqrt{15}} = 6.827 \pm .865$, i.e. $5.962,\ 7.692$ lb. and we say that population mean lie between 5.962 lb. and 7.692 lb.

Test for difference of means

When n_1 and n_2 the sizes of the sample are small and the population variances σ_1^2 and σ_2^2 are unknown, and assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

A test for $H_0 : \mu_1 - \mu_2 = 0$ is then given by the statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \xi_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

where $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$.

For $H_A : \xi_0 \neq 0$, reject H_0, if

$$|t| > t_{\alpha/2,(n_1+n_2-2)\,df}.$$

Note: Following the usual procedure, it can be found that the confidence limits to $\mu_1 - \mu_2$ are

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,(n_1+n_2-2)}\, s_p \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}.$$

Remark:

i) This $t-$statistic with df $(n_1 + n_2 - 2)$ may be applied even when σ_1^2 and σ_2^2 are unequal, provided the ratio σ_1^2/σ_2^2 does not differ from unity by more than 0.4. If it does, requiring the use of alternative procedures. Various solutions have been offered for this problem called the **Fisher-Behrens problem**.

ii) When the sizes of the samples are same say equal to n $(n_1 = n_2 = n)$ and data are paired *e.g.* we want to test the effect of a particular drug, for inducing sleep. Let x_i and y_i $(i = 1, 2, \cdots, n)$ be the readings in hrs of sleep on the $i-$th individual before and after the drug is given respectively. Hence we use paired $t-$test instead of difference of the mean test.

In such problems, the difference between the paired observations are considered as a random sample (which is assumed from normal population) as:

$$d_i = x_i - y_i, \quad (i = 1, 2, \cdots, n)$$

$$H_0 : \mu_x = \mu_y \qquad \text{(i.e. drug is not effective)}$$

$$H_A : \mu_y > \mu_x \qquad \text{(i.e. drug is effective)}$$

Under H_0, test statistic will be

$$t = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1}, \text{ where } \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

H_0 is to be rejected on the basis of the given sample if $t > t_{\alpha,n-1}$.

Note: If n is large, then $t-$distribution cannot be used. We use a corresponding test based on

$$Z = \frac{\bar{d}}{\sigma/\sqrt{n}} \sim N(0,1).$$

If σ is unknown then $\hat{\sigma}^2 = s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$.

Exercise 3: For a random sample of 10 cows fed on diet A, the increase in weights in pounds in a certain period were

10, 6, 16, 17, 13, 12, 8, 14, 15, 9.

For another random sample of 12 cows fed on diet B, the increase in weights in pounds in the same period were

7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17.

Test whether diets A and B differ significantly as regard the effect on increase in weights.

Solution: Mean gain in weights on diet A

$$\bar{x}_1 = \frac{120}{10} = 12 \text{ lbs.}$$

Mean gain in weights on diet B

$$\bar{x}_2 = 15 \text{ lbs.}$$

and $s_1^2 = \frac{1}{n_1 - 1}\sum_{\alpha=1}^{n_1}(x_{1\alpha} - \bar{x}_1)^2 \quad \Rightarrow (n_1 - 1)s_1^2 = \sum_{\alpha=1}^{n_1}(x_{1\alpha} - \bar{x}_1)^2 = 120$

Similarly,

$$(n_2 - 1)s_2^2 = \sum_{\alpha=1}^{n_2}(x_{2\alpha} - \bar{x}_2)^2 = 314, \text{ thus}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 21.7.$$

The hypothesis of interest is

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0$$

under H_0, the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}.$$

Substitute the known values in the test statistic, then

$$t = \frac{12 - 15}{4.65\sqrt{\frac{1}{10} + \frac{1}{12}}} = -1.6.$$

From the table $t_{0.025,20} = 2.086$.

Since $|t| = 1.6 < t_{0.025,\,20} = 2.086$ and hence we accept H_0 i.e. the difference between two means is not significance (two diets do not differ significantly as regards their effect on increase in weights).

TEST BASED ON CHI-SQUARE DISTRIBUTION

The square of a standard normal variate is known as a chi-square variate with one degree of freedom (df), i.e.

$$X \sim N(\mu, \sigma^2), \text{ then } Z = \left(\frac{X-\mu}{\sigma}\right) \sim N(0,1)$$

and $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi_1^2$.

In general,

If $X_1, X_2, \cdots, X_n$ are n independent normal variates with mean μ_i and variance σ_i^2, i.e.,

$$X_i \sim N(\mu_i, \sigma_i^2), \text{ then } \quad Z_i = \left(\frac{X_i-\mu_i}{\sigma_i}\right) \sim N(0,1)$$

and $\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2 \sim \chi_n^2$.

Also,

If $X_1, X_2, \cdots, X_n$ be a random sample from a normal population with mean μ and variance σ^2, then

$$\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \sim \chi_{n-1}^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{\sigma^2}(n-1)S^2$$

$$\Rightarrow \quad \frac{n-1}{\sigma^2}S^2 \sim \chi_{n-1}^2.$$

Thus the sampling distribution of variance in sampling from a normal population is χ_{n-1}^2.

Note: The pdf of χ^2-distribution with n df is given by

$$f(\chi^2) = \frac{1}{2^{n/2}\Gamma(n/2)}(\chi^2)^{(n/2)-1}\exp(\chi^2/2), \qquad 0 \le \chi^2 \le \infty.$$

Characteristics of chi-square

i) The mean and variance of χ^2 depends on its degree of freedom, i.e.

mean $= n$ and variance $= 2n$, for n degree of freedom

ii) If $\chi^2/2 = x \quad \Rightarrow \quad 2x = \chi^2$ and $d\chi^2 = 2dx$, then the pdf of x

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)}(2x)^{(n/2)-1}e^{-x} = \frac{1}{\Gamma(n/2)}x^{(n/2)-1}e^{-x}, \quad x \ge 0.$$

Which is the pdf of gamma variate with parameter $n/2$.

iii) Since the probability function $f(\chi^2)$ does not involve any population parameter, χ^2-distribution is some times considered to be a non-parametric distribution.

Conditions for validity of χ^2-distribution

i) The sample observations should be independent and normally distributed.

ii) Constraints imposed upon the observations if any, should be linear.

iii) No estimated or theoretical cell frequency should be less than 5.

iv) The total frequency should be reasonably large, say more than 50

Application of χ^2-distribution

i) Application to Genetics, where in breeding experiments it is desired to determine whether observed results are consistent with expected segregations.

ii) Application to contingency table, where it is used in testing independence in a contingency table.

iii) Application to the fitting of distribution, it is used as test of goodness of fit of a distribution.

iv) To test the variance of a normal population.

v) To test the homogeneity of independent estimates of population variance.

Test for population variance

Case I: μ known and σ^2 unknown.

Let X_i is a normal variate with mean μ and variance σ^2, i.e. $X_i \sim N(\mu,\sigma^2)$, then

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 \sim \chi_n^2 .$$

Suppose that we want to test the null hypothesis

$$H_0: \quad \sigma^2 = \sigma_0^2$$

$$H_A: \text{ i) } \quad \sigma^2 > \sigma_0^2, \text{ ii) } \sigma^2 < \sigma_0^2, \text{ iii) } \sigma^2 \neq \sigma_0^2$$

Under H_0, the test statistic becomes

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma_0}\right)^2 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu)^2 .$$

H_0 is rejected under $H_A: \sigma^2 > \sigma_0^2$, if for the given sample $\chi^2 > \chi^2_{\alpha,n}$ (This $\chi^2_{\alpha,n}$ is the tabulated value at $\alpha\%$ level of significance).

For $H_A: \sigma^2 < \sigma_0^2$, H_0 is rejected if for the given sample $\chi^2 < \chi^2_{1-\alpha,n}$.

For $H_A: \sigma^2 \neq \sigma_0^2$, H_0 is rejected if for the given sample $\chi^2 < \chi^2_{1-\alpha/2,n}$ or $\chi^2 > \chi^2_{\alpha/2,n}$.

Case II: both μ and σ^2 unknown.

Suppose that we want to test the null hypothesis

$$H_0: \quad \sigma^2 = \sigma_0^2$$

$$H_A: \text{ i) } \quad \sigma^2 > \sigma_0^2, \text{ ii) } \sigma^2 < \sigma_0^2, \text{ iii) } \sigma^2 \neq \sigma_0^2$$

Under H_0, the test statistic will be

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_0}\right)^2 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \sim \chi_{n-1}^2.$$

The value of χ^2, computed from the given sample, is compared with $\chi_{\alpha, n-1}^2$ or $\chi_{1-\alpha, n-1}^2$ according as the

$$H_A: \ \sigma^2 > \sigma_0^2 \quad \text{or} \quad \sigma^2 < \sigma_0^2.$$

For $H_A: \sigma^2 \neq \sigma_0^2$, the computed value is to be compared with both

$$\chi_{1-\alpha/2, n-1}^2 \quad \text{and} \quad \chi_{\alpha/2, n-1}^2,$$

H_0 being rejected if the value of χ^2 is less than former or greater than the latter value.

Remark: If the sample size n is large (> 30), then, we use an approximated test statistic as given

$$\sqrt{2\chi^2} \sim N(\sqrt{2n-1}, 1), \text{ so that}$$

$$Z = (\sqrt{2\chi^2} - \sqrt{2n-1}) \sim N(0,1). \text{ Here } n \text{ is a degree of freedom.}$$

Exercise 1) A firm manufacturing rivets want to limit variations in their length as much as possible. The lengths (in cm) of 10 rivets manufactured by a new process are 2.15 1.99 2.05 2.12 2.17 2.01 1.98 2.03 2.25 1.93. In the past, the variation in length of rivets manufactured by the firm has been 0.021 cm, examine whether the new process seems to be superior to the old.

Solution:

$$H_0: \ \sigma^2 = 0.021 \text{ cm}$$

$$H_A: \ \sigma^2 < 0.021 \text{ cm}$$

Test statistic, under H_0

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_0}\right)^2 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{0.021} \times 0.091 = 4.326.$$

Since $\chi^2 = 4.326 > \chi_{0.99,\ 9}^2 = 2.088$ at 1% level and $\chi^2 = 4.326 > \chi_{0.95,\ 9}^2 = 3.325$ at 5% level. Thus H_0 is to be accepted and we conclude that the new process does not seem to be superior to the old.

Exercise 2) Test $H_0: \sigma^2 = 100$ versus $H_A: \sigma^2 \neq 100$, given that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 225$ for a random sample of size 50 from a normal population.

Solution: Here, $n = 50$, $S^2 = 225$.

Test statistic, under H_0

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma_0}\right)^2 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}$$

$$= \frac{(n-1)S^2}{\sigma_0^2}.$$

Since n is large, the test statistic under H_0 will be

$$Z = (\sqrt{2\chi^2} - \sqrt{2n-1}) = \sqrt{\frac{2\times 49\times 225}{100}} - \sqrt{99} = 5.05.$$

Since $|Z| > 3$, it is significant at all level of significance and hence H_0 is rejected and we conclude that $\sigma^2 \neq 100$.

Confidence limit for population Variance

We know that

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Thus, by definition

$$\Pr\left[\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right] = 1-\alpha$$

or $$\Pr\left[\frac{\chi^2_{1-\alpha/2,n-1}}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{\alpha/2,n-1}}{(n-1)S^2}\right] = 1-\alpha$$

or $$\Pr\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right] = 1-\alpha.$$

Exercise 3) A random sample of size 15 gives an unbiased estimate of sample variance equal to 1.05 cm^2. Find the 90% confidence interval for σ^2.

Solution: We know that

$$\Pr\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \frac{1}{\sigma^2} \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right] = 1-\alpha,$$ the $100(1-\alpha)\%$ confidence limit for σ^2.

Therefore, 90% limit for σ^2 will be

$$\Pr\left[\frac{14\times 1.05}{23.685} \le \sigma^2 \le \frac{14\times 1.05}{6.571}\right] = 0.90$$

and 95% limit for σ^2 will be

$$\Pr\left[\frac{14\times 1.05}{26.119} \le \sigma^2 \le \frac{14\times 1.05}{5.629}\right] = 0.95.$$

Test for goodness of fit

Let a sample of size n be drawn from a population in which P is the proportion of members having the character A and Q is the proportion of members having the character not$-A$. Suppose x be the number of members in the sample having the character A and $n-x$ be the number of members having the character not$-A$. (These x and $n-x$ are also called observed frequencies) and the expected frequencies of these observed frequencies will be

$$E(x) = nP \text{ and } E(n-x) = nQ.$$

Then, the deviation of the observed frequencies from the theoretical frequencies or expected frequencies will be

$$x - nP \text{ and } n - x - nQ = n - x - n(1-P) = nP - x.$$

In this situation, we use the test statistic for testing the hypothesis as given

$$Z = \frac{x-nP}{\sqrt{nPQ}} \sim N(0,1), \text{ so that}$$

$$Z^2 = \frac{(x-nP)^2}{nPQ} = \frac{(x-nP)^2}{nP} + \frac{(n-x-nQ)^2}{nQ} \tag{1}$$

Consider RHS of equation (1)

$$\frac{(x-nP)^2}{nP} + \frac{(n-x-nQ)^2}{nQ} = \frac{Q(x-nP)^2 + P(n-x-nQ)^2}{nPQ}$$

$$= \frac{Q(x^2 + n^2P^2 - 2nxP) + P(nP-x)^2}{nPQ}$$

$$= \frac{(1-P)(x^2 + n^2P^2 - 2nxP) + P(n^2P^2 + x^2 - 2nxP)}{nPQ}$$

$$= \frac{x^2 + n^2P^2 - 2nxP}{nPQ} = \frac{(x-nP)^2}{nPQ}.$$

For large samples, Z is an asymptotic standard normal variate, so that Z^2 is a χ^2 variate with 1 degree of freedom.

If O_1 and O_2 represent the observed frequencies in the two classes and E_1 and E_2 the expected frequencies then equation (1) can be written as

$$Z^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \sim \chi_1^2.$$

Suppose we have k classes instead of two, with $O_1, O_2, \cdots, O_k$ as the observed frequencies and $E_1, E_2, \cdots, E_k$ as the expected frequencies (with $\sum O_i = \sum E_i$), we form the statistic

$$Z^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1} .$$

This test is used in such situation when one wishes to decide if an observed distribution of frequencies is incompatible with some preconceived or hypothesised distribution, e.g. one may wish to determine whether or not a sample of observed values of some random variable is compatible with the hypothesis that it was drawn from a population values that is normally distributed or Poisson or binomial or any other distribution. (a given set of data approximate a given distribution say normal).

H_0: The population from which the sample was drawn is normally distributed.

H_A: The sampled population is not normally distributed.

Case I: Hypothetical population completely specified.

Under H_0, the test statistic is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1} ,$$

where O_i is a set of observed frequencies and E_i is the corresponding set of expected (hypothetical) frequencies (with $\sum O_i = \sum E_i$).

When $\chi^2 < \chi^2_{\alpha, k-1}$, we accept our null hypothesis, that the sample came from normally distributed population.

Case II: Hypothetical population unspecified.

In this situation the parameters of the hypothetical distribution are unknown which may be estimated from given sample, are considered. However, the number of degree will get reduced as

$$(k-1) - (\text{number of parameters estimated say } 'r') = k - 1 - r .$$

Thus, the test statistic will be as follows:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1-r} .$$

When $\chi^2 < \chi^2_{\alpha, k-1-r}$, we accept our H_0, that the sample came from normally distributed population.

Remarks:

i) All expected frequencies should be no less than 5. In case, the expected frequencies is less than 5 adjacent categories may be combined to achieve the suggested minimum expected frequency, combining reduces the number of categories and therefore, the number of degree of freedom. E.g. suppose first three categories having expected frequencies less than 5, than combine these three as one, such a way here we are losing two categories as well as degree of freedom.

ii) Cochran suggested that if relatively few expected frequencies are less than 5 (say just one out of five or two or more out of ten or more) than even as low a value as 1 is allowable for an expected frequency in using χ^2.

Exercise 4) Given below the number of male and female in 1000 families having 5 children

Number of male births	0	1	2	3	4	5
Number of female births	5	4	3	2	1	0
Number of family	40	300	250	200	180	30

Is this result consistent with the hypothesis that male and female births are equally probable?

Solution:

H_0: Given data are consistent with the hypothesis of equal probability for male and female births.

Under H_0

$p =$ the probability of male births $=$ the probability of female births $= 1/2 = q$

and the probability of $'x' =$ number of male births in a family of 5

$$p(x) = \binom{5}{x} p^x q^{5-x}, \quad x = 0,1,\cdots,5.$$

$$= \binom{5}{x}\left(\frac{1}{2}\right)^5.$$

The expected frequencies of x male births is given by

$f(x) = N\,p(x)$, where N is the total frequencies.

$$= 1000\binom{5}{x}\left(\frac{1}{2}\right)^5 = 31.25\binom{5}{x}.$$

Substituting $x = 0,1,2,\cdots,5$, we get the expected frequencies as follows:

$$f(0) = 31.25\binom{5}{0} = 31.25 \cong 31.$$

$$f(1) = 31.25\binom{5}{1} = 31.25\times 5 = 156.25 \cong 156.$$

$$f(2) = 31.25\binom{5}{2} = 31.25\times 10 = 312.5 \cong 313.$$

$$f(3) = 31.25\binom{5}{3} = 31.25\times 10 = 312.5 \cong 313.$$

$$f(4) = 31.25\binom{5}{4} = 31.25\times 5 = 156.25 \cong 156.$$

$$f(5) = 31.25\binom{5}{5} = 31.25 \cong 31.$$

Test statistic will be

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= \left[\frac{(40-31.25)^2}{31.25} + \frac{(300-156.25)^2}{156.25} + \cdots + \frac{(30-31.25)^2}{31.25} \right] = 191.36 .$$

Tabulated $\chi^2_{0.05,5} = 11.07$.

Since $\chi^2 = 191.96 > \chi^2_{0.05,5} = 11.07$ at 5% level of significance, H_0 is rejected and conclude that equal probability for male and female births may not be accepted.

CHAPTER VI

DESIGN OF EXPERIMENT

Design of experiment means how to design an experiment in the sense that how the observations or measurements should be obtained to answer a query in a valid, efficient and economical way. The designing of the experiment and the analysis of obtained data are inseparable. If the experiment is designed properly keeping in mind the question, then the data generated is valid and proper analysis of data provides the valid statistical inferences. If the experiment is not well designed, the validity of the statistical inferences is questionable and may be invalid. It is important to understand first the basic terminologies used in the experimental design.

Experiment: An experiment is a procedure which is done to make a discovery, test a hypothesis, to understand cause and effect relationships. In other words, a way of getting an answer to a question which the experimenter wants to know.

Experiment Unit: For conducting an experiment, the experimental material is divided into smaller parts and each part is referred to as experimental unit. It is that unit to which a treatment is applied. Examples are manufactured item, plot of land, etc.

Treatment: An object whose effect is measured and compared is called treatment. For example, drugs, fertilizers, teaching methods, etc.

Experimental Error: It describes the failure of two identically treated experimental units to give identical results. For example, mistake in data entry, systematic error or mistakes in the design of the experiment itself or random error, caused by environmental condition or other unpredictable factor.

Basic Principles of a Good Experimental Design

There are three basic principles of getting a good experimental design.

Randomization: The principle of randomization involves the allocation of treatment to experimental units at random to avoid any bias in the experiment resulting from the influence of some extraneous unknown factor (such as rising temperature, drifting, calibration of instruments, etc.) that may affect the experiment.

Replication: The repetition of treatment by applying them to more than one experimental unit is known as replication. In any experimental situation replication is necessary in order to get an estimate of the experimental error variation caused due to uncontrollable factors. Replication provides an efficient way of increasing the precision of an experiment. The precision increases with the increase in the number of observations.

Local Control (error control): The entire experimental material is suppose to be heterogeneous, may be divided into different groups by taking homogeneous units together and then treatment, may be allocated randomly to different units in each group. This process is known as local control. For example, if the experimental units are divided into different groups (blocks) such that they are homogeneous with in block, then the variation among the blocks is eliminated and ideally, the error component will contain the variation due to the treatment only. This will, in turn, increase the efficiency.

Completely Randomized Design (CRD)

A completely randomized (CR) design, which is the simplest type of the basic designs, may be defined as a design in which the treatments are assigned to experimental units completely

at random. That is, the randomization is done without any restrictions. The design is completely flexible, i.e., any number of treatments and any number of units per treatment may be used. Moreover, the number of units per treatment need not be equal. A completely randomized design is considered to be most useful in situations where

i) the experimental units are homogeneous

ii) the experiments are small such as laboratory experiments and greenhouse studies, and

iii) some experimental units are likely to be destroyed or fail to respond.

Layout of CRD

CRD is the one in which all the experimental units are taken in a single group which are homogenous as for as possible.

The randomization procedure for allotting the treatments to various units will be as follows.

Step 1: Determine the total number of experimental units.

Step 2: Assign a plot number to each experimental units starting from left to right for all rows.

Step 3: Assign the treatments to the experimental units by using random numbers.

The statistical model for CRD with one observation per unit

$Y_{ij} = \mu + t_i + e_{ij}$, where

$Y_{ij} =$ value of the variate in the $j-$th replication of the $i-$th treatment; $i = 1, 2, \cdots, t$; $j = 1, 2, \cdots, r_i$

$\mu =$ over all mean effect

$t_i =$ true effect of the $i-$th treatment

$e_{ij} =$ error term of the $j-$th unit receiving $i-$th treatment.

The arrangement of data in CRD is as follows:

	Treatments						
	T_1	T_2	$\cdots$	T_i	$\cdots$	T_k	
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{11}	
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{12}	
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	
	y_{1r_1}	y_{2r_1}	$\cdots$	y_{ir_i}	$\cdots$	y_{1r_1}	
Total	Y_1	Y_2	$\cdots$	Y_i	$\cdots$	Y_k	GT

The null hypothesis will be

$H_0 : T_1 = T_2 = \cdots = T_t$, (there is no significant difference between the treatments)

and the alternative hypothesis is

$H_A : T_i$ are not equal (there is significant difference between the treatments)

The different steps in forming the analysis of variance table for a CRD are:

Correction Factor $(CF) = \frac{(GT)^2}{n}$, where, n is the total number of observations.

Total sum of square $(TSS) = \sum_{i=1}^{k}\sum_{j=1}^{r_i} y_{ij}^2 - CF$

Treatment sum of square $(SST) = \frac{Y_1^2}{r_1} + \frac{Y_2^2}{r_2} + \cdots + \frac{Y_k^2}{r_k} - CF = \sum_{i=1}^{k}\frac{Y_i^2}{r_i} - CF$

Error sum of square $(SSE) = \sum_{i=1}^{k}\sum_{j=1}^{r_i} y_{ij}^2 - \sum_{i=1}^{k}\frac{Y_i^2}{r_i} = TSS - SST$

Form the following ANOVA table and calculate F value

Source of variation	$d.f.$	SS	MSS	Variance ratio F
Treatments	$t-1$	SST	$MST = \frac{SST}{t-1}$	$\frac{MST}{MSE}$
Error (within treatment)	$n-t$	SSE	$MSE = \frac{SSE}{n-t}$	
Total	$n-1$	TSS		

Compare the calculated F with the critical value of $F_{\alpha;(t-1),(n-t)}$ (corresponding to treatment degree of freedom and error degree of freedom) so that acceptance or rejection of the null hypothesis can be determined.

If null hypothesis is rejected that indicates there is significance differences between the different treatments.

Calculate critical difference $(CD) = SE(d) \times t_{\alpha;(n-t)}$, where

$$SE(d) = \sqrt{MSE\left(\frac{1}{r_i} + \frac{1}{r_j}\right)} \text{ and}$$

r_i = number of replications for treatment i

r_j = number of replications for treatment j

$t_{\alpha;(n-t)}$ is the critical t value for error degree of freedom at specified level of significance, either 5% or 1%.

Advantages of a CRD

i) Its lay out is very easy.

ii) There is complete flexibility in this design i.e. any number of treatments and replications for each treatment can be tried.

iii) Whole experimental material can be utilized in this design.

iv) This design yields maximum degrees of freedom for experimental error.

v) The analysis of data is simplest as compared to any other design.

vi) Even if some values are missing the analysis can be done.

Disadvantages of a CRD

i) It is difficult to find homogeneous experimental units in all respects and hence CRD is seldom suitable for field experiments as compared to other experimental designs.

ii) It is less accurate than other designs.

Example: The following table gives the yield in kgs per plot of five varieties of wheat after being applied to each of four plots in a completely randomized design.

Varieties	Yield in kgs				Total	Treatment means
A	8	8	6	10	32 (T_1)	8 (T_1)
B	10	12	13	9	44 (T_2)	11 (T_2)
C	18	17	13	16	64 (T_3)	16 (T_3)
D	12	10	15	11	48 (T_4)	12 (T_4)
E	8	11	9	8	36 (T_5)	9 (T_5)
Grand Total					224	

Analysis

Correction Factor $(CF) = \dfrac{(GT)^2}{n} = \dfrac{(224)^2}{4 \times 5} = 2508.8$

Total sum of square $(TSS) = \sum_{i=1}^{k}\sum_{j=1}^{r_i} y_{ij}^2 - CF = 8^2 + 8^2 + \cdots + 9^2 + 8^2 - 2508.8 = 207.2$

Treatment sum of square $(SST) = \dfrac{T_1^2 + T_2^2 + T_3^2 + T_4^2 + T_5^2}{r} - CF$

$$= \frac{(32)^2 + (44)^2 + (64)^2 + (48)^2 + (36)^2}{4} - 2508.8 = 155.2$$

Error sum of square $(SSE) = TSS - SST = 207.2 - 155.2 = 52.0$.

Table for Analysis of Variance

Source of variation	$d.f.$	SS	MSS	Variance ratio F
Between Varieties (Treatments)	4	155.2	38.8	11.8
Within Varieties (Error)	15	52.0	3.47	
Total	19	207.2		

Here $F-$test indicates that there are significant difference between the variety means since the observed value of the variance ratio is significant at 5% level of significance. Now we wish to know as to which variety is the best and also which variety show the significant difference among themselves. This can be done with the help of critical difference (CD).

Now, standard error of the difference between two treatment means is

$$SE(d) = \sqrt{MSE\left(\frac{1}{r_i} + \frac{1}{r_j}\right)} = \sqrt{\frac{2 \times MSE}{r}} = \sqrt{\frac{2 \times 3.47}{4}} = 1.32 .$$

Therefore,

Critical Difference $(CD) = SE(d) \times t_{0.05;(15)} = 1.32 \times 2.131 = 2.81$.

Since the $F-$test in the analysis of variance indicates significant differences between the varieties, we are justified in comparing the varieties with the help of the above Critical Difference value.

Varieties	A	B	C	D	E	CD
Mean yields	8	11	16	12	9	2.81

Conclusion represent symbolically (bar chart notation):

The varieties can be compared by setting them in the descending (or ascending) order of their mean yields in the following manner.

$$C \qquad \underline{D \qquad B} \qquad \overline{\qquad E} \qquad \underline{\qquad A}$$

The varieties which do not differ significantly have been underlined by a bar. This method of underlining the treatments which do not differ significantly is the way of indicating the significance and non-signification of individual comparisons.

Exercise

1) The following table gives the yields of five varieties of paddy with four replications each by using completely randomized Design.

Varieties	Yield in kg.			
A	8	8	6	10
B	10	12	13	9
C	18	17	13	16
D	12	10	15	11
E	8	11	9	8

Analyze the date and draw your conclusions.

2) Below are given the plan and yield in kg per plot of a completely randomized design for testing the effect of five different fertilizers A, B, C, D and E.

D 20	E 17	B 21	E 16	D 15
A 8	C 17	A 9	D 13	B 17
B 12	D 19	E 18	C 18	A 15
C 16	A 8	C 18	A 10	E 15
E 13	B 16	D 23	B 14	C 19

Analyze the data and state your conclusions.

Randomized Block Design (RBD)

In order to control the variability in one direction in the experimental material, it is desirable to divide the experimental units into homogeneous groups of units known as blocks. The treatments are randomly allocated separately to each unit of these blocks. This procedure gives rise to a design known as Randomized Block Design.

OR

It can be defined as an arrangement of t treatment in r blocks such that each treatment occurs precisely once in each block.

Layout of RBD

Suppose we have four treatments A, B, C and D each replicated four times. We divide the whole experimental area into four relatively homogeneous blocks and each block into four units or plots. Treatments are allocated at random to the units of the blocks, fresh randomization being done for each block. A particular layout may be as follows:

Block 1	A	B	D	C
Block 2	D	C	B	A
Block 3	C	B	A	D
Block 4	A	D	C	B

Analysis

For analysis of this design we use the following linear additive model

$$Y_{ij} = \mu + t_i + b_j + e_{ij}, \; i = 1, 2, \cdots, t; \; j = 1, 2, \cdots, r,$$

where

$\mu =$ over all mean effect

$t_i =$ true effect of the $i-$th treatment

$b_j =$ true effect of the $j-$th block

$e_{ij} =$ error due to random component.

The results of RBD can be arranged in a two way table according to the replications (blocks) and treatments.

There will be $r \times t$ observations in total, where r stands for number of replications and t for number of treatments. .

The data are arranged in a two way table form by representing treatments in rows and replications (block) in columns.

Treatment	Replication						Total
	1	2	…	j	…	r	
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1r}	T_1
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2r}	T_2
$\vdots$	$\vdots$	$\vdots$	…	$\vdots$	…	$\vdots$	$\vdots$

i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{ir}	T_i
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
t	y_{t1}	y_{t2}	$\cdots$	y_{tj}	$\cdots$	y_{tr}	T_t
Total	B_1	B_2	$\cdots$	B_j	$\cdots$	B_r	GT

Let

$$Y_{i.} = \sum_{j=1}^{r} y_{ij} = T_i = \text{Total of } i-\text{th treatment}$$

$$Y_{.j} = \sum_{i=1}^{t} y_{ij} = B_j = \text{Total of } j-\text{th block}$$

$$Y_{..} = \sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij} = GT = \sum_{i=1}^{t} T_i = \sum_{j=1}^{r} B_j \ .$$

The following formulae are used for the analysis of variance of this design

Correction Factor $(CF) = \dfrac{(GT)^2}{n}$, where, n is the total number of observations.

Total sum of square $(TSS) = \sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}^2 - CF$

Treatment sum of square $(SST) = \sum_{i=1}^{k} \dfrac{T_i^2}{r} - CF = \dfrac{1}{r}\sum_{i=1}^{t} Y_{i.}^2 - CF$

Block sum of square $(SSB) = \sum_{j=1}^{r} \dfrac{B_j^2}{t} - CF = \dfrac{1}{t}\sum_{j=1}^{r} Y_{.j}^2 - CF$

Error sum of square $(SSE) = TSS - SST - SSB$

The null hypothesis will be

$H_0 : T_1 = T_2 = \cdots = T_t$, (there is no significant difference between the treatments)

$B_1 = B_2 = \cdots = B_r$, (there is no significant difference between the blocks)

and the alternative hypothesis is

$H_A : T_i$ are not equal (there is significant difference between the treatments)

B_j are not equal (there is significant difference between the blocks)

Form the following ANOVA table and calculate F value

Source of variation	$d.f.$	SS	MSS	Variance ratio F
Treatments	$t-1$	SST	$MST = \frac{SST}{t-1}$	$\frac{MST}{MSE} = F$
Replication (block)	$r-1$	SSB	$MSB = \frac{SSB}{r-1}$	$\frac{MSB}{MSE} = F$
Error (within treatment)	$(r-1)(t-1)$	SSE	$MSE = \frac{SSE}{(r-1)(t-1)}$	
Total	$rt-1 = n-1$	TSS		

Compare the calculated F with the critical values of $F_{\alpha;(t-1),(r-1)(t-1)}$ and $F_{\alpha;(r-1),(r-1)(t-1)}$, respectively, so that acceptance or rejection of the null hypothesis can be determined.

If null hypothesis is rejected that indicates there is significance differences.

Calculate critical difference $(CD) = SE(d) \times t_{\alpha;(r-1)(t-1)}$, where

$$SE(d) = \sqrt{\frac{2\,MSE}{r}}.$$

Advantages of a RBD

The precision is more in RBD. The amount of information obtained in RBD is more as compared to CRD. RBD is more flexible. Statistical analysis is simple and easy.

Even if some values are missing, still the analysis can be done by using missing plot technique.

Disadvantages of RBD

When the number of treatments is increased, the block size will increase. If the block size is large maintaining homogeneity is difficult and hence when more number of treatments is present this design may not be suitable.

Analysis of RBD with one missing observation

Suppose the observation y_{11} is missing in the layout of RBD, denoted this y_{11} by y

Treatment	Replication (block)						Total
	1	2	…	j	…	r	
1	y	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1r}	T_1'
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2r}	T_2
$\vdots$	$\vdots$	$\vdots$	…	$\vdots$	…	$\vdots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{ir}	T_i
$\vdots$	$\vdots$	$\vdots$	…	$\vdots$	…	$\vdots$	$\vdots$
t	y_{t1}	y_{t2}	$\cdots$	y_{tj}	$\cdots$	y_{tr}	T_t
Total	B_1'	B_2	…	B_j	…	B_r	G'

T_1': Total of $(r-1)$ values of first treatment

B_1': Total of $(t-1)$ values of first block

G': Total of $(rt-1)$ values

Total sum of square $(TSS) = y^2 + \text{terms not containing } y - \dfrac{(G'+y)^2}{n}$

Treatment sum of square $(SST) = \dfrac{(T_1'+y)^2}{r} + \text{terms not containing } y - \dfrac{(G'+y)^2}{n}$

Block sum of square $(SSB) = \dfrac{(B_1'+y)^2}{t} + \text{terms not containing } y - \dfrac{(G'+y)^2}{n}$

Error sum of square $(SSE) = TSS - SST - SSB$

$$= y^2 - \frac{(T_1' + y)^2}{r} - \frac{(B_1' + y)^2}{t} + \frac{(G' + y)^2}{n} + \text{ terms not containing } y .$$

Now the best value of y is that minimizes SSE. So the method of least square the least value of y is given by differentiating one with respect to y equating to zero.

$$\frac{\partial SSE}{\partial y} = 0 = 2y - 2\frac{(T_1' + y)}{r} - 2\frac{(B_1' + y)}{t} + 2\frac{(G' + y)}{rt} + 0$$

$$y - \frac{(T_1' + y)}{r} - \frac{(B_1' + y)}{t} + \frac{(G' + y)}{rt} = 0$$

$$\left(1 - \frac{1}{r} - \frac{1}{t} + \frac{1}{rt}\right) y = \frac{T_1'}{r} + \frac{B_1'}{t} - \frac{G'}{rt}$$

$$\left(\frac{rt - t - r + 1}{rt}\right) y = \frac{tT_1' + rB_1' - G'}{rt}$$

$$y = \frac{tT_1' + rB_1' - G'}{rt - t - r + 1} = \frac{tT_1' + rB_1' - G'}{(r-1)(t-1)} .$$

This value of y is put in missing place. All SS are computed in the analysis, are computed as usual. However the degree of freedom in the total and error SS both are reduced by one.

The adjusted treatment is obtained by subtracting the so called adjusted factor $\frac{(tT_1' + rB_1' - G')^2}{t(t-1)(r-1)^2}$ from the treatment SS.

Example: The yields of six nitrogen treatments on a crop in kgs along with the plan of the experiment are given below. The number of blocks is five and the nitrogen treatments have been represented by A, B, C, D, E and F.

Block I	Block II	Block III	Block IV	Block V
D 17	B 12	E 23	A 28	F 75
C 12	C 15	A 30	F 64	C 14
F 70	E 26	C 16	B 9	D 20
B 6	A 26	D 20	D 23	B 7
A 20	D 10	F 56	E 33	E 30
E 28	F 62	B 10	C 14	A 23

It is required to analyze the data.

Analysis

The first step in the analysis of data is to tabulate yield figures according to block and treatments in the following manner.

Varieties	Blocks					Total	Treatment means
	I	II	III	IV	V		
A	20	26	30	28	23	127 (T_1)	25.4
B	9	12	10	9	7	47 (T_2)	9.4
C	12	15	16	14	14	71 (T_3)	14.2
D	17	10	20	23	20	90 (T_4)	18.0
E	28	26	23	35	30	142 (T_5)	28.4
F	70	62	56	64	75	327 (T_6)	65.4
Total	156 (B_1)	156 (B_2)	156 (B_3)	156 (B_4)	156 (B_5)	156 (GT)	

Correction Factor $(CF) = \dfrac{(GT)^2}{n} = \dfrac{(804)^2}{5\times 6} = 21547.2$.

Total sum of square $(TSS) = \sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}^2 - CF$

$$= (20)^2 + (9)^2 + \cdots + (75)^2 - 21547.2 = 10466.8.$$

Treatment sum of square $(SST) = \sum_{i=1}^{k} \dfrac{T_i^2}{r} - CF = \dfrac{1}{r}\sum_{i=1}^{t} Y_{i.}^2 - CF$

$$= \frac{(127)^2 + (47)^2 + \cdots + (327)^2}{5} - 21547.2 = 10167.2.$$

Block sum of square $(SSB) = \sum_{j=1}^{r} \dfrac{B_j^2}{t} - CF = \dfrac{1}{t}\sum_{j=1}^{r} Y_{.j}^2 - CF$

$$= \frac{(156)^2 + (151)^2 + \cdots + (169)^2}{6} - 21547.2 = 61.4.$$

Error sum of square $(SSE) = TSS - SST - SSB = 10466.8 - 10167.2 - 61.4 = 418.2$.

Table for Analysis of Variance

Source of variation	$d.f.$	SS	MSS	Variance ratio F
Between Varieties (Treatments)	5	10167.2	2033.44	97.24
Blocks	4	61.4	15.35	
Within Varieties (Error)	20	418.2	20.19	
Total	29	10646.8		

It is clear from the table that this observed value of F is significant at 5% level of significance which proves that there are significant differences between the treatment means. Now, we have to test the significance of the difference between the individual treatments, and this will be done with the help of CD as usual.

Critical Difference

S.E. of the difference between any two treatment means is

$$SE(d) = \sqrt{\frac{2 \times MSE}{r}} = \sqrt{\frac{2 \times 20.91}{5}} = 2.89.$$

Therefore,

Critical Difference $(CD) = SE(d) \times t_{0.05} = 2.89 \times 2.086 = 6.03$.

Conclusions represented symbolically

The treatments have been compared by setting them in the descending order of their mean yields in the following manner

Varieties	F	E	A	D	C	B
Mean yield	65.4	28.4	25.4	18.0	14.2	9.4

The treatments which do not differ significantly have been underlined by a bar. The treatment 'F" has been found to be the best.

Exercise

1) The yield of rice (in kg) with five fertilizers tested in four blocks using RBD is given the following layout. Analysis the data and interpret your conclusion.

Block 1	Block 2	Block 3	Block 4
B 10	C 13	A 19	D 20
C 16	A 21	D 24	E 36
A 20	D 21	E 32	B 9
D 23	E 31	B 10	C 13
E 33	B 11	C 14	A 24

2) An experiment was conducted in RBD to study to comparative performance of yield of six varieties of oranges (kg/plot) are given below. Analyze the data and give your conclusion.

Treatment	Blocks				
	B_1	B_2	B_3	B_4	B_5
V_1	5.5	5.9	6.3	6.5	6.7
V_2	7.4	7.7	7.9	7.5	8.1
V_3	4.6	5.1	5.3	4.9	4.7
V_4	5.0	5.8	5.6	6.1	5.3
V_5	6.7	6.2	6.9	6.8	6.0
V_6	8.2	7.9	7.5	7.2	6.9

Latin Square Design

To control two way heterogeneity (variability) in the experimental material, we use the design known as Latin Square Design (LSD). In this design two restrictions are imposed by forming blocks in two directions, row wise and column wise. Treatments are allocated in such a way that every treatment occurs once and only once in each row and in each column.

A LSD of size m is an arrangement of m Latin letters into m^2 positions such that every row and every column contain every treatment precisely once.

Layout of LSD

In this design, the whole experimental material is divided into m^2 experimental units and arranged in a square so that each row as well as each column contains m units. The m treatments are then allocated at random to these rows and columns in such a way that every treatment occurs once and only once in each row and each column.

In LSD the treatments are usually denoted by A B C D etc. For a 5×5 LSD the arrangements may be

A	B	C	D	E
B	A	E	C	D
C	D	A	E	B
D	E	B	A	C
E	C	D	B	A
Square 1				

A	B	C	D	E
B	A	D	E	C
C	E	A	B	D
D	C	E	A	B
E	D	B	C	A
Square 2				

A	B	C	D	E
B	C	D	E	A
C	D	E	A	B
D	E	A	B	C
E	A	B	C	D
Square 3				

Analysis of LSD

For analysis of this design, we use the linear additive model

$$Y_{ijk} = \mu + r_i + c_j + t_k + e_{ijk},\ i, j, k = 1, 2, \cdots, t$$

where

Y_{ijk} is the observation on the $k-$th treatment in the $i-$th row and $j-$th column

μ is the general mean effect

r_i is the effect due to $i-$th row

c_j is the effect due to $j-$th column

t_k is the effect due to $k-$th treatment

e_{ijk} is error due to random component.

Let these be t treatments arranged in t rows and t columns.

The following formulae are used for the analysis of variance of this design

Correction Factor $(CF) = \dfrac{(GT)^2}{t^2}$, where, n is the total number of observations.

Total sum of square $(TSS) = \sum_{i=1}^{t}\sum_{j=1}^{t}\sum_{k=1}^{t} y_{ijk}^2 - CF$

Treatment sum of square $(SST) = \sum_{k=1}^{t} \dfrac{T_k^2}{t} - CF$

Row sum of square $(SSR) = \sum_{i=1}^{t} \dfrac{R_i^2}{t} - CF$

Column sum of square $(SSC) = \sum_{j=1}^{t} \dfrac{C_j^2}{t} - CF$

Error sum of square $(SSE) = TSS - SST - SSR - SSC$

The null hypothesis will be

$H_0 : T_1 = T_2 = \cdots = T_t$, (there is no significant difference between the treatments)

$R_1 = R_2 = \cdots = R_t$, (there is no significant difference between the rows)

$C_1 = C_2 = \cdots = C_t$, (there is no significant difference between the column)

Form the following ANOVA table and calculate F value

Source of variation	$d.f.$	SS	MSS	Variance ratio F
Treatments	$t-1$	SST	$MST = \frac{SST}{t-1}$	$\frac{MST}{MSE} = F$
Rows	$t-1$	SSR	$MSR = \frac{SSR}{t-1}$	$\frac{MSR}{MSE} = F$
Columns	$t-1$	SSC	$MSC = \frac{SSC}{t-1}$	$\frac{MSC}{MSE} = F$
Error (within treatment)	$(t-1)(t-2)$	SSE	$MSE = \frac{SSE}{(t-1)(t-2)}$	
Total	t^2-1	TSS		

If the calculated value of F is greater than tabulated value of $F_{\alpha;(t-1)(t-2)}$, then reject H_0, otherwise accept it.

Advantages of LSD

LSD is more efficient than RBD or CRD. This is because of double grouping that will result in small experimental error.

When missing values are present, missing plot technique can be used and analyzed.

Disadvantages of LSD

This design is not as flexible as RBD or CRD as the number of treatments is limited to the number of rows and columns. LSD is seldom used when the number of treatments is more than 12. LSD is not suitable for treatment less than 5.

Analysis of LSD with one missing observation

Suppose the observation occurring in the first row and first column and receiving the first treatment is missing, denoted this by y

Rows	Columns						Total
	C_1	C_2	$\cdots$	C_j	$\cdots$	C_t	
R_1	y		$\cdots$		$\cdots$		$R_1' + y$
R_2			$\cdots$		$\cdots$		R_2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
R_i			$\cdots$		$\cdots$		R_i
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
R_t			$\cdots$		$\cdots$		R_t
Total	$C_1' + y$	C_2	$\cdots$	C_j	$\cdots$	C_t	$G' + y$

Total sum of square $(TSS) = y^2 +$ terms not containing $y - \dfrac{(G'+y)^2}{t^2}$

Treatment sum of square $(SST) = \dfrac{(T_1'+y)^2}{t} +$ terms not containing $y - \dfrac{(G'+y)^2}{t^2}$

Row sum of square $(SSR) = \dfrac{(R_1'+y)^2}{t} +$ terms not containing $y - \dfrac{(G'+y)^2}{t^2}$

Column sum of square $(SSC) = \dfrac{(C_1'+y)^2}{t} +$ terms not containing $y - \dfrac{(G'+y)^2}{t^2}$

Error sum of square $(SSE) = TSS - SST - SSR - SSC$

$$= y^2 - \frac{(T_1'+y)^2}{t} - \frac{(R_1'+y)^2}{t} - \frac{(C_1'+y)^2}{t} + 2\frac{(G'+y)^2}{t^2} + \text{ terms not containing } y.$$

Now the best value of y is that minimizes SSE. So the method of least square the least value of y is given by differentiating one with respect to y equating to zero.

$$\frac{\partial SSE}{\partial y} = 0 = 2y - 2\frac{(T_1' + y)}{t} - 2\frac{(R_1' + y)}{t} - 2\frac{(C_1' + y)}{t} + 4\frac{(G' + y)}{t^2} + 0$$

$$y - \frac{(T_1' + y)}{t} - \frac{(R_1' + y)}{t} - \frac{(C_1' + y)}{t} + 2\frac{(G' + y)}{t^2} = 0$$

$$\frac{t^2 y - t(T_1' + y) - t(R_1' + y) - t(C_1' + y) + 2(G' + y)}{t^2} = 0$$

$$(t^2 - 3t + 2)\, y = t(T_1' + R_1' + C'_1) - 2G'$$

$$y = \frac{t\ (T_1' + R_1' + C'_1) - 2G'}{(t^2 - 3t + 2)}.$$

The value of y is substituted for the missing place and analysis of the data is carried out as usual. Substract one degree of freedom from error and total.

Example: Below are given the plan and yield in kgs/plot of a 5×5 Latin square experiment on the wheat crop carried out for testing the effects of five, manorial treatments A, B, C, D, and E. 'A' denotes control.

B 15	A 8	E 17	D 20	C 17	$R_1 = 77$
A 9	D 21	C 19	E 16	B 13	$R_2 = 78$
C 18	B 12	D 23	A 8	E 17	$R_3 = 78$
E 18	C 16	A 10	B 15	D 23	$R_4 = 82$
D 22	E 15	B 13	C 18	A 10	$R_5 = 78$

$C_1 = 82$, $C_2 = 72$, $C_3 = 82$, $C_4 = 77$, $C_5 = 80$; $GT = 393$

Analyze the data and state your conclusions.

Analysis

Correction Factor $(CF) = \dfrac{(GT)^2}{t^2} = \dfrac{(397)^2}{25} = 6177.96$

Total sum of square $(TSS) = \sum_{i=1}^{t}\sum_{j=1}^{t}\sum_{k=1}^{t} y_{ijk}^2 - CF = 152 + 82 + \cdots + 102 - CF$

$$= 6661 - 6177.96 = 483.04$$

Treatment sum of square $(SST) = \sum_{k=1}^{t} \frac{T_k^2}{t} - CF = \frac{(45)^2 + (68)^2 + \cdots + (83)^2}{5} - 6177.96$

$= 454.64$

Row sum of square $(SSR) = \sum_{i=1}^{t} \frac{R_i^2}{t} - CF = \frac{(77)^2 + (78)^2 + \cdots + (78)^2}{5} - 6177.96 = 3.04$

Column sum of square $(SSC) = \sum_{j=1}^{t} \frac{C_j^2}{t} - CF = \frac{(77)^2 + (78)^2 + \cdots + (78)^2}{5} - 6177.96 = 3.04$

Error sum of square $(SSE) = TSS - SST - SSR - SSC = 483.04 - 454.64 - 3.04 - 14.24$

$= 11.12$

Source of variation	$d.f.$	SS	MSS	Variance ratio F
Treatments	4	454.24	113.66	123.34
Rows	4	3.04	0.76	
Columns	4	14.24	3.56	
Error (within treatment)	12	11.12	0.92	
Total	24	483.04		

The observed highly significant value of the variance ratio indicates that there are significant differences between the treatment means.

Critical Difference

S.E. of the difference between the treatment means

$$SE(d) = \sqrt{\frac{2 \times MSE}{r}} = \sqrt{\frac{2 \times 0.92}{5}} = 0.61.$$

Therefore,

Critical Difference $(CD) = SE(d) \times t_{0.05} = 0.61 \times 2.179 = 1.33$.

Summary of Results

Treatment means will be calculated from the original table on totals.

Treatments	A	B	C	D	E	CD 5%A
Mean yield	14.2	18.0	17.6	21.8	25.4	1.33

Conclusions represented symbolically

The treatments have been compared by setting them in the descending order of their mean yields in the following manner

Treatments	D	C	E	B	A
Mean yield	21.8	17.6	25.4	18.0	14.2

The treatment D is the best of all. The treatments C and E do not differ significantly each other.

The yields obtained by applying every one of the manorial treatment is significantly higher that obtained without applying any manure.

Exercise

1) An oil company tested four different blends of gasoline for fuel efficiency according to a Latin square design in order to control for the variability of four different drivers and four different models of cars. Fuel efficiency was measured in miles per gallon (mpg) after driving cars over a standard course.

Fuel Efficiencies (mpg) For 4 Blends of Gasoline

(Latin Square Design: Blends Indicated by Letters A-D)

Deriver	Car Model			
	I	II	III	IV
1	D 15	B 33	C 13.2	A 29.1
2	B 16.3	C 26.6	A 19.4	D 22.8
3	C 10.8	A 31.1	D 17.1	B 30.3
4	A 14.7	D 34.0	B 19.7	C 21.6

Analyze the data and draw your conclusion.

2) The numbers of wireworms counted in the plots of Latin square following soil fumigations (L, M, N, O, P) in the previous year were

Rows	Columns				
1	P 4	O 2	N 5	L 1	M 3
2	M 5	L 1	O 6	N 5	P 3
3	O 4	M 8	L 1	P 5	N 4
4	N 12	P 7	M 7	O 10	L 5
5	L 5	N 4	P 3	M 6	O 9

Analyze the data and draw your conclusions.

3) The following layout presents the observations made on 5 treatments A, B, C, D and E in an experiment of paddy crop by adopting LSD. The figures indicate the grain yield of paddy in kg/plot. Analyze the data and draw your conclusion.

Rows	Columns				
	B 5	D 6	E 3	A 10	C 12
	C 9	A 4	B 6	E 5	D 5
	D 8	C 15	A 7	B 6	E 5
	E 5	B 8	C 13	D 9	A 5
	A 9	E 6	D 12	C 16	B 8

Factorial Experiments

When two or more number of factors are investigated simultaneously in a single experiment such experiments are called as factorial experiments. For example, the yield of a crop depends on particular variety of a crop being used and also on particular fertilizer applied.

Terminologies

Factor: Factor refers to a set of related treatments. We may apply of different doses of nitrogen to a crop. Hence nitrogen irrespective of doses is a factor.

Levels of a factor: Different states or components making up a factor are known as the levels of that factor, e.g. different doses of nitrogen.

Types of factorial Experiment

A factorial experiment is named based on the number of factors and levels of factors. For example, when there are 3 factors each at 2 levels the experiment is known as $2\times2\times2$ or 2^3 factorial experiments.

If there are 2 factors each at 3 levels then it is known as 3×3 or 3^2 factorial experiment.

In general if there are n factors each with p levels then it is known as p^n factorial experiment.

For varying number of levels the arrangement is described by the product. For example, an experiment with 3 factors each at 2 levels, 3 levels and 4 levels respectively, then it is known as $2\times3\times4$ factorial experiment.

If all the factors have the same number of levels the experiment is known as symmetrical factorial otherwise it is called as mixed factorial.

Factors are represented by capital letters. Treatment combinations are usually by small letters.

Simple and Main Effects

Simple effect of a factor is the difference between its responses for a fixed level of other factors.

Main effect is defined as the average of the simple effects.

Interaction is defined as the dependence of factors in their responses. Interaction is measured as the mean of the differences between simple effects.

Advantages

1. In such type of experiments we study the individual effects of each factor and their interactions.

2. In factorial experiments a wide range of factor combinations are used.

3. Factorial approach will result in considerable saving of the experimental

resources, experimental material and time.

Disadvantages

1. When number of factors or levels of factors or both are increased, the number of treatment combinations increases. Consequently block size increases. If block size increases it may be difficult to maintain homogeneity of experimental material.

This will lead to increase in experimental error and loss of precision in the experiment.

2. All treatment combinations are to be included for the experiment irrespective of its importance and hence this results in wastage of experimental material and time.

3. When many treatment combinations are included the execution of the experiment and statistical analysis become difficult.

2^2 Factorial Experiment

Consider two factors A and B, each at two levels , say a_0, a_1 and b_0, b_1, so there will be $2 \times 2 = 4$ treatment combination. They are enumerated as follows:

$a_0\, b_0$ or (1) : A and B both at the first level

$a_1\, b_0$ or a : A at second level and B at first level

$a_0\, b_1$ or b : A at first level and B at second level

$a_1\, b_1$ or $a\,b$: A and B both at the second level

Note: The first level of A and B is generally expressed by the absence of the corresponding letter in the treatment combination.

These four treatment combinations may be compared using a CRD or RBD or LSD.

In 2^2 experiment in RBD, the analysis will be the same as stated in RBD with the number of treatment $t = 4$, and the analysis of 2^2 experiment in LSD will be the same as stated in LSD with $t = 4$.

In factorial experiment, one is more interested in the separate tests about main effects and interactions, which are performed by splitting the Treatment SS carrying 3 degree of freedom in to three orthogonal components, each carrying a single degree of freedom and each associated with main effect or an interaction.

Consider $Yate's$ method to obtain the factorial effects and their SS from the treatment totals which is as follows:

Treatment combination	Treatment total	I	II	Effect total
(1)	[1]	$[1]+[a]=u_1$	u_1+u_2	G
a	$[a]$	$[b]+[a\,b]=u_2$	u_3+u_4	$[A]$
b	$[b]$	$[a]-[1]=u_3$	u_2-u_1	$[B]$
ab	$[a\,b]$	$[a\,b]-[b]=u_4$	u_4-u_3	$[AB]$

Setup the hypothesis

H_0 : i) The blocks as well as treatments are homogeneous

ii) All the main effects and interactions are significant

Correction Factor $(CF) = \dfrac{(G)^2}{4r}$

Sum of square due to replication (block) $(SSB) = \dfrac{B_i^2}{4} - CF$

Total sum of square $(TSS) = \sum_{i=1}^{r}\sum_{j=1}^{r} y_{ij}^2 - CF$

Sum of square due to any main effect or the interaction effect is obtained as follows;

Sum of square due to main effect A $(SSA) = \frac{[A]^2}{2^2 r}$

Sum of square due to main effect B $(SSB) = \frac{[B]^2}{2^2 r}$

Sum of square due to interaction effect AB $(SSAB) = \frac{[AB]^2}{2^2 r}$

Error sum of square $(SSE) = TSS - SSR - SSA - SSB - SSAB$

Form the following ANOVA table and calculate F value

Source of variation	$d.f.$	SS	MSS	Variance ratio F
Replication (Block)	$r-1$	SSB	$MSB = \frac{SSB}{r-1}$	$\frac{MSB}{MSE} = F$
Main effect A	1	SSA	$MSA = SSA$	$\frac{MSA}{MSE} = F$
Main effect B	1	SSB	$MSB = SSB$	$\frac{MSB}{MSE} = F$
Interaction effect AB	1	$SS\ AB$	$MS\ AB = SS\ AB$	$\frac{MSAB}{MSE} = F$
Error (within treatment)	$3(r-1)$	SSE	$MSE = \frac{SSE}{3(r-1)}$	
Total	$2^2 r - 1$	TSS		

If calculated value is greater than tabulated value at given level of significance, then reject H_0, otherwise accept it.

Example: An experiment was planned to study the effects of sulphate of potash and super phosphate on the yield of potatoes. All the combinations of 2 levels of super phosphate [0% (p_0) and 5% (p_1)/acre] and 2 levels sulphate of potash [0% (k_0) and 5% (k_1)/acre] were studied in a randomized block design with 4 replications for each. The following yield were obtained: (Gupta and Kapoor, 2000)

Block				
I	(1) 23	k 25	p 22	kp 38
II	p 40	(1) 26	k 36	kp 38
III	(1) 29	k 20	kp 30	p 20
IV	kp 34	k 31	p 24	(1) 28

Analyse the data and give your conclusions.

Solution: Taking deviation from $y = 29$, we rearrange the given table in the following table for computations.

Block	I	II	III	IV	Treatment Totals T_i	T_i^2
(1)	-6	-3	0	-1	-10	100
k	-4	7	-9	2	-4	16
p	-7	11	-9	-5	-10	100
kp	9	9	1	5	24	576
Block totals B_i	-8	24	-17	1	$G = 0$	
B_i^2	64	576	289	1		

H_0 : i) The blocks as well as treatments are homogeneous

ii) All the main effects and interactions are significant

Correction Factor $(CF) = \dfrac{(G)^2}{4r} = \dfrac{0}{16} = 0$

Total sum of square $(TSS) = \sum_{i=1}^{r}\sum_{j=1}^{r} y_{ij}^2 - CF = 660 - 0 = 660$

Sum of square due to replication (block) $(SSB) = \dfrac{1}{4}\sum_{i=1}^{r} B_i^2 - CF = \dfrac{64+576+289+1}{4} = 232.5$

Sum of square due to treatment $(SST) = \dfrac{1}{4}\sum_{i=1}^{r} T_i^2 - CF = \dfrac{100+16+100+576}{4} = 198$

Error sum of square $(SSE) = 660 - 232.5 - 198.0 = 229.5$.

We now compute the factorial effect totals by Yates method

Treatment combination	Treatment total	I	II	Effect total
(1)	-10	$-10-4=-14=u_1$	u_1+u_2	$G=0$
k	-4	$-10+24=14=u_2$	u_3+u_4	$40=[K]$
p	-10	$-4+10=6=u_3$	u_2-u_1	$28=[P]$
kp	24	$24+10=34=u_4$	u_4-u_3	$28=[KP]$

Sum of square due to any main effect or the interaction effect is obtained as follows;

Sum of square due to main effect K $(SSK) = \dfrac{[K]^2}{2^2 r} = \dfrac{(40)^2}{16} = 100$

Sum of square due to main effect P $(SSP) = \frac{[P]^2}{2^2 r} = \frac{(28)^2}{16} = 49$

Sum of square due to interaction effect KP $(SSKP) = \frac{[KP]^2}{2^2 r} = \frac{(28)^2}{16} = 49$

Form the following ANOVA table and calculate F value

Source of variation	$d.f.$	SS	MSS	Variance ratio F
Replication (Block)	3	$SSB = 232.5$	$MSB = \frac{SSB}{r-1} = 77.5$	$\frac{MSB}{MSE} = F = 3.04$
Treatments	3	$SST = 198.0$	$MST = \frac{SST}{r-1} = 66.0$	$\frac{MST}{MSE} = F = 2.59$
Main effect K	1	$SSK = 100$	$MSK = SSK = 100$	$\frac{MSK}{MSE} = F = 3.92$
Main effect P	1	$SSP = 49$	$MSP = SSP = 49$	$\frac{MSP}{MSE} = F = 1.92$
Interaction effect KP	1	$SS\,KP = 49$	$MS\,KP = SS\,KP = 49$	$\frac{MSKP}{MSE} = F = 1.92$
Error (within treat.)	9	$SSE = 229.5$	$MSE = \frac{SSE}{3(r-1)} = 25.5$	
Total	15	$TSS = 660.0$		

As in each of the cases, the computed value of F is less than the corresponding theoretical value, there are no significant main or interaction effects presented in the experiment. The blocks as well as treatments do not differ significantly.

Remark: It may be noted that

$SSK + SSP + SS\,KP = 100 + 49 + 49 = 198 = SST$, as it should be.

2^3 Factorial Experiment

Consider three factors say A, B and C, each at two levels , say (a_0, a_1), (b_0, b_1) and (c_0, c_1), respectively, so that there are $2 \times 2 \times 2 = 8$ treatment combinations and this combination can be written in a systematic way as

$a_0\, b_0\, c_0$ or (1) : A, B and C all at the first level

$a_1\, b_0\, c_0$ or a : A at second level and B, C at first level

$a_0\, b_1\, c_0$ or b : A, C at first level and B at second level

$a_0\, b_0\, c_1$ or c : A, B at first level and C at second level

$a_1\, b_1\, c_0$ or ab : C at first level and A, B at second level

$a_1\, b_0\, c_1$ or ac : B at first level and A, C at second level

$a_0\, b_1\, c_1$ or bc : A at first level and B, C at second level

$a_1\, b_1\, c_1$ or abc : A, B and C all at the second level

2^3 factorial experiment can be performed as a CRD with 8 treatments. RBD with r replications, each replicates containing 8 treatments or LSD with $t = 8$ and data can be analyzed accordingly.

In 2^3 experiment we split up the Treatment SS with 7 degree of freedom into 7 orthogonal components corresponding to the three main effects A, B and C, three first order (on two factor) interactions AB, AC, BC and one second order (on three factors) interaction ABC each carrying one degree of freedom.

To obtain the factorial effects and their SS from the treatment totals, consider $Yate's$ method which is as follows:

Treatment combination	Treatment total	I	II	Effect total
(1)	[1]	$[1]+[a]=u_1$	$u_1+u_2=v_1$	$v_1+v_2=G$
a	$[a]$	$[b]+[ab]=u_2$	$u_3+u_4=v_2$	$v_3+v_4=[A]$
b	$[b]$	$[c]+[ac]=u_3$	$u_5+u_6=v_3$	$v_5+v_6=[B]$
ab	$[ab]$	$[bc]+[abc]=u_4$	$u_7+-u_8=v_4$	$v_7+v_8=[AB]$
c	$[c]$	$[a]-[1]=u_5$	$u_2-u_1=v_5$	$v_2-v_1=[C]$
ac	$[ac]$	$[ab]-[b]=u_6$	$u_4-u_3=v_6$	$v_4-v_3=[AC]$
bc	$[bc]$	$[ac]-[c]=u_7$	$u_6-u_5=v_7$	$v_6-v_5=[BC]$
abc	$[abc]$	$[abc]-[bc]=u_8$	$u_8-u_7=v_8$	$v_8-v_7=[ABC]$

Correction Factor $(CF)=\dfrac{(G)^2}{8r}$

Sum of square due to replication (block) $(SSB)=\dfrac{1}{8}\sum_i B_i^2-CF$, where B_i is the total of $i-$th block

Total sum of square $(TSS)=\sum_{i=1}^{r}\sum_{j=1}^{r}y_{ij}^2-CF$

Sum of square due to any main effect or the interaction effect is obtained as follows;

Sum of square due to main effect A $(SSA)=\dfrac{[A]^2}{8r}$

Sum of square due to main effect B $(SSB)=\dfrac{[B]^2}{8r}$

Sum of square due to main effect C $(SSC)=\dfrac{[C]^2}{8r}$

Sum of square due to AB $(SSAB)=\dfrac{[AB]^2}{8r}$

Sum of square due to AC $(SSAC)=\dfrac{[AC]^2}{8r}$

Sum of square due to BC $(SSBC) = \dfrac{[BC]^2}{8r}$

Sum of square due to ABC $(SSABC) = \dfrac{[ABC]^2}{8r}$

Error sum of square $(SSE) = TSS - SSR - SSA - SSB - SSAB - \cdots - SSABC$

Form the following ANOVA table and calculate F value

Source of variation	$d.f.$	SS	MSS	Variance ratio F
Replication (Block)	$r-1$	SSB	$MSB = \frac{SSB}{r-1}$	$\frac{MSB}{MSE} = F$
Main effect A	1	SSA	$MSA = SSA$	$\frac{MSA}{MSE} = F$
Main effect B	1	SSB	$MSB = SSB$	$\frac{MSB}{MSE} = F$
Main effect C	1	SSC	$MSC = SSC$	$\frac{MSAB}{MSE} = F$
Interaction effect AB	1	$SS\ AB$	$MS\ AB = SS\ AB$	
Interaction effect AC	1	$SS\ AC$	$MS\ AC = SS\ AC$	
Interaction effect BC	1	$SS\ BC$	$MS\ BC = SS\ BC$	
Interaction effect ABC	1	$SS\ ABC$	$MS\ ABC = SS\ ABC$	
Error (within treatment)	$7(r-1)$	SSE	$MSE = \frac{SSE}{3(r-1)}$	
Total	$8r-1$	TSS		

If calculated value is greater than tabulated value at given level of significance, then reject H_0, otherwise accept it.

Example: The following table gives the layout and the results of a 2^3 factorial design laid out 4 replicates. The purpose of the experiment is to determine the effect of different kinds of fertilizers Nitrogen N, Potash K and Phosphate P on potato crop yield.

Block I				Block II			
nk 291	kp 391	p 312	np 373	kp 407	p 324	k 272	nk 306
(1) 101	k 265	n 106	nkp 450	n 89	nkp 449	np 338	(1) 106
Block III				**Block IV**			
p 323	(1) 87	np 324	kp 423	np 361	nk 272	n 103	p 324
nk 334	k 279	n 128	nkp 471	k 302	(1) 131	nkp 437	kp 435

Solution. H_0 : Blocks as well as treatments are homogeneous.

Block effects are eliminated by carrying out the analysis of the given design as an RBD for 8 treatment combinations and 4 blocks. The initial calculations are therefore as follows:

Total number of observations $= 4 \times 8 = 32$

The block totals, the treatment total and the grand total are summarized in the following table:

Block Totals	1 = 2289	Treatment Totals	(1) = 425
	2 = 2291		$n = 426$
	3 = 2369		$k = 1118$
	4 = 2375		$nk = 1203$
			$p = 1283$
			$np = 1396$
			$kp = 1666$
			$nkp = 1807$

Grand total $= G = 9324$

Correction Factor $(CF) = \dfrac{(G)^2}{32} = 2716780.5$.

Total sum of square $(TSS) = \sum_{i=1}^{r}\sum_{j=1}^{r} y_{ij}^2 - CF$

$$= [(291)^2 + (391)^2 + \cdots + (437)^2 + (445)^2] - 2716780.5$$

$$= 465337.5$$

Sum of square due to replication (block) $(SSB) = \dfrac{1}{8}\sum_i B_i^2 - CF$

$$= \frac{21740988}{8} - 2716780.5 = 843.0$$

Sum of square due to treatment $(SST) = \dfrac{1}{4}\sum_{i=1}^{r} T_i^2 - CF = \dfrac{12694944}{4} - 2716780.5 = 456955.5$.

Error sum of square $(SSE) = 465337.5 - 456955.5 - 843.0 = 7539$.

We shall now break up the treatment sum of square with 7 degree of freedom. For this we use Yates method for finding the various factorial effect totals and their sum of square.

Treatment combination	Treatment total	I	II	Effect total
(1)	$[1] = 425$	$[1] + [n] = 851 = u_1$	$u_1 + u_2 = 3172 = v_1$	$v_1 + v_2 = 9324 = G$
n	$[n] = 426$	$[k] + [nk] = 2321 = u_2$	$u_3 + u_4 = 6152 = v_2$	$v_3 + v_4 = 340 = [N]$
k	$[k] = 1118$	$[p] + [np] = 2679 = u_3$	$u_5 + u_6 = 86 = v_3$	$v_5 + v_6 = 2264 = [K]$

nk	$[nk]=1203$	$[kp]+[nkp]=3473=u_4$	$u_7+u_8=254=v_4$	$v_7+v_8=112=[NK]$
p	$[p]=1283$	$[n]-[1]=1=u_5$	$u_2-u_1=1470=v_5$	$v_2-v_1=2980=[P]$
np	$[np]=1396$	$[nk]-[k]=85=u_6$	$u_4-u_3=794=v_6$	$v_4-v_3=168=[NP]$
kp	$[kp]=1666$	$[np]-[p]=113=u_7$	$u_6-u_5=84=v_7$	$v_6-v_5=-676=[KP]$
nkp	$[nkp]=1807$	$[nkp]-[kp]=141=u_8$	$u_8-u_7=28=v_8$	$v_8-v_7=-56=[NKP]$

Sum of square due to any main effect or the interaction effect is obtained as follows;

Sum of square due to main effect N $(SSN)=\dfrac{[N]^2}{2^3r}=\dfrac{(340)^2}{32}=3612.5$

Sum of square due to main effect K $(SSK)=\dfrac{[K]^2}{2^3r}=\dfrac{(2264)^2}{32}=160178.0$

Sum of square due to main effect P $(SSP)=\dfrac{[P]^2}{2^3r}=\dfrac{(2980)^2}{32}=277512.5$

Sum of square due to interaction effect NK $(SS\ NP)=\dfrac{[NK]^2}{32}=\dfrac{(112)^2}{32}=392.0$

Sum of square due to interaction effect NP $(SS\ NP)=\dfrac{[NP]^2}{32}=\dfrac{(168)^2}{32}=882.0$

Sum of square due to interaction effect KP $(SS\ KP)=\dfrac{[KP]^2}{32}=\dfrac{(-676)^2}{32}=14280.5$

Sum of square due to interaction effect nKP $(SS\ NKP)=\dfrac{[NKP]^2}{32}=\dfrac{(-56)^2}{32}=98.0$

Form the following ANOVA table and calculate F value

Source of variation	$d.f.$	SS	MSS	Variance ratio F
Replication (Block)	3	$SSB=843.0$	$MSB=\frac{SSB}{r-1}=281.0$	$\frac{MSB}{MSE}=F=0.78$
Treatments	7	$SST=456955.5$	$MST=65729.35$	$F=181.83$
Main effect N	1	$SSN=3612.5$	$MSK=3612.5$	$F=10.06$
Main effect K	1	$SSK=160178.0$	$MSK=160178.0$	$F=446.1$
Main effect P	1	$SSP=277512.5$	$MSP=277512.5$	$F=773.01$
Interaction effect NK	1	$SS\ NK=392.0$	$MS\ NK=392.0$	$F=1.09$
Interaction effect NP	1	$SS\ NP=882.0$	$MS\ NP=882.0$	$F=2.45$
Interaction effect KP	1	$SS\ KP=14280.5$	$MS\ KP=14280.5$	$F=39.7$
Interaction effect NKP	1	$SS\ NKP=98.0$	$MS\ NKP=98.0$	$F=0.27$
Error	21	$SSE=7539.0$	$MSE=359.0$	
Total	15	$TSS=465337.5$		

From the ANOVA table we find that

i) Replicates or blocks are homogeneous.

ii) Treatments differ significantly. Among the factorial effects, all the main effects N, K and P, and the interaction KP are significant.

BASIC CONCEPT OF SAMPLE SURVEYS

A sample survey is a method of drawing an inference about the characteristics of a population or universe by observing a part of the population. For example, when one has to make an inference about a large lot and is not practicable to examine each individual member of the lot, one always takes help of sample surveys, that is to say one examines only a few member of the lot and, on the basis of this sample information, one makes decisions about the whole lot. Thus, a person wanting to purchase a basket of oranges may examine a few oranges from the basket and on that basis make his decision about the whole basket.

Such methods are extensively used by government bodies throughout the world for assessing, different characteristics of national economy as are required for taking decisions regarding the impositions of taxes, fixation of prices and minimum wages etc. and for planning and projection of future economic structure, for estimation of yield rates and acreages under different crops, number of unemployed persons in the labour forces, construction of cost of living indices for persons in different professions and so on.

Sample survey techniques are extensively used in market research surveys for assessing the preferential pattern of consumers for different types of products, the potential demand for a new product which a company wishes to introduce, scope for any diversification in the production schedule, and so on.

Thus, sampling may become unavoidable because we may have limited resources in terms of money and / or man hours, or it may be preferred because of practical convenience.

Sampling is first broadly classified as **Subjective** and **Objective**.

Any type of sampling which depends upon the personal judgment or discretion of the sampler himself is called **Subjective**. But the sampling method which is fixed by a sampling rule or is independent of the sampler's own judgment is **Objective sampling**.

Objective Sampling

Non-probabilistic | Probabilistic and mixed

In non-probabilistic objective sampling, there is a fixed sampling rule but there is no probability attached to the mode of selection, e.g. selecting every $5-$th individual from a list. If, however, the selection of the first individual is made in such a manner that each of the first 10 gets an equal chance of being selected, it becomes a case of mixed sampling, if for each individual there is a definite pre-assigned probability of being selected, the sampling is said to be probabilistic.

Elementary unit or simply unit: It is an element or a group of elements, on which observations can be made or from which the required statistical information can be ascertained according to a well defined procedure, examples of unit are person, family, household, farm, factory, tree, a period of time such as an hour, day etc.

Population: The collection of all units of a specified type in a given region at a particular point or a period of time is termed as a population or inverse. For example, a population of persons, families, farms, cattle, houses or automobiles in a region or a population of trees or a birds in a forest etc.

A population is said to be finite population or an infinite population according to as the number of units in it is finite or infinite.

Sampling units: Elementary units or groups of such units, which, besides being clearly defined, identifiable and observable, are convenient for purposes of sampling, are called sampling units. For example, in a family budget enquiry, usually a family is considered as a sampling unit, since it is formed to be convenient for sampling for ascertaining the required information. In a crop survey, a farm or a group of farms owned or operated by a household may be considered as the sampling units.

Sampling frame: For using sampling methods in the collection of data, it is essential to have a frame of all the sampling units belonging to the population to be studied with their proper identification particulars and such a frame is called the sampling frame. This may be a list of units with their identification particulars.

As the sampling frame forms the basic material from which a sample is drawn, it should be insured that the frame contains all the sampling units of the population under consideration but excludes units of any other population.

Sample: A sample is a subset of a population selected to obtain information concerning the characteristics of the population. In other words, one or more sampling units selected from a population according to some specified procedure are said to constitute a sample.

Random sample: A random or probability sample is a sample drawn in such a manner that each unit in the population has a predetermined probability of selection.

Estimator: An estimator is a statistic obtained by a specified procedure for estimating a population parameter. The estimator is a random variable, as its value differs from sample to sample and the samples are selected with specified probabilities.

The particular value, which the estimator takes for a given sample, is known as an estimate.

The difference between the estimator (t) and the parameter (θ) is called **error**.

An estimator (t) is said to be unbiased estimator for the parameter (θ) if, $E(t)=\theta$, otherwise biased. Thus bias is given by

$$E(t-\theta)=B(t)$$

The mean of squares of error taken from θ is called **mean square error** (MSE). Mathematically it is defined as

$$MSE(t)=E(t-\theta)^2.$$

The MSE may be considered to be a measure of **accuracy** with which the estimator t estimates the parameter θ.

The expected value of the squared deviation of the estimator from its expected value is termed **sampling variance**. It is a measure of the divergence of the estimator from its expected value and is given by

$$V(t)=E[t-E(t)]^2.$$

This measure of variability may be termed the **precision of the estimator** t.

The relation between MSE and sampling variance or between accuracy and precision can be obtained as

$$MSE(t)=E(t-\theta)^2=E[t-E(t)+E(t)-\theta]^2$$

$$=E[t-E(t)]^2+[E(t)-\theta]^2=V(t)+[B(t)]^2, \text{ since } E[t-E(t)]=0.$$

This shows that MSE of t is the sum of the sampling variance and the square of the bias. However, if t is an unbiased estimator of θ, the MSE and sampling variance are the same.

The square root of the sampling variance is termed as **standard error** of the estimator t.

The ratio of the standard error of the estimator to the expected value of the estimator is known as **relative standard error** or the coefficient of variation of the estimator.

Sample space: The collection of all possible sample, sequence, sets is called the sample space.

Sampling design: The combination of the sample space and the associated probability measure is called a sampling design. For example, let $N=4$, $n=2$ and the probability of selection for different samples is

Sample	$(1,2)$	$(1,3)$	$(1,4)$	$(2,3)$	$(2,4)$	$(3,4)$
Probability	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

The above table gives the sampling design.

Sampling and complete enumeration

The total count of all units of the population for a certain characteristics is known as complete enumeration, also termed census survey. The money, man-power and time required for carrying out complete enumeration will generally be large and there are many situations with limited means where complete enumeration will not be possible, where recourse to selection of a few units will be helpful. When only a part, called sample, is selected from the population and examine, it is called sample enumeration or sample survey.

A sample survey will usually be less expensive then a census survey and the desired information will obtain in less time. This does not imply that economy is the only consideration in conducting a sample survey. It is most important that a degree of accuracy of results is also maintained. Occasionally, the technique of sample survey is applied to verify that the results obtained from the census surveys. The main advantages or merits of sample survey over census survey may be outlined as follows:

i) Reduced cost of survey,

ii) Greater speed of getting results,

iii) Greater accuracy of results,

iv) Greater scope, and

v) Adaptability

Sample survey has its own limitations and the advantages of sampling over complete enumeration can be derived only if

i) the units are drawn in a scientific manner

ii) an appropriate sampling technique is used, and

iii) the size of units selected in the sample is adequate.

Basic principles of sample surveys

Two basic principles for sample surveys are

i) Validity

ii) Optimization

The principle of optimization takes into account the factors of

a) Efficiency

b) Cost

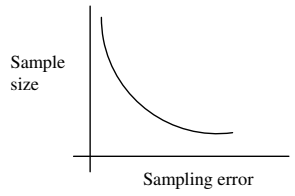
By validity, we mean that the sample should be so selected that the results could be interpreted objectively in terms of probability. The principle will be satisfied by selecting a probability sample, which ensures that there is some definite, pre-assigned probability for each individual of the population.

Efficiency is measured by the inverse of the sample variance of the estimator.

Cost is measured by the expenditure incurred in terms of money or man-hours. The principle of optimization insures that a given level of efficiency will be reached with minimum cost or that the maximum possible efficiency will be attained with a given level of cost.

Sampling and non-sampling errors

The error which arises due to only a sample (a part of population) being used to estimate the population parameters and draw inferences about the population is termed **sampling error** or **sampling fluctuation**. Whatever may be the degree of cautiousness in selecting a sample; there will always be a difference between the parameter and its corresponding estimate. This error is inherent and unavoidable in any and every sampling scheme. A sample with the smallest sampling error will always be considered a good representative of the population. This error can be reduced by increasing the size of the sample (number of units selected in the sample). In fact, the decrease in sampling error is inversely proportional to the square root of the sample size and the relationship can be examined graphically as below:



When the sample survey becomes a census survey, the sampling error becomes zero.

Non-sampling error

The non-sampling errors primarily arise at the following stages:

i) Failure to measure some of units in the selected sample

ii) Observational errors due to defective measurement technique

iii) Errors introduced in editing, coding and tabulating the results.

Non-sampling errors are present in both the complete enumeration survey and the sample survey. In practice, the census survey results may suffer from non-sampling errors although these may be free from sampling error. The non-sampling error is likely to increase with increase in sample size, while sampling error decreases with increase in sample size.

SIMPLE RANDOM SAMPLING

A procedure for selecting a sample of size n out of a finite population of size N in which each of the possible distinct samples has an equal chance of being selected is called **random sampling or simple random sampling.**

We may have two distinct types of simple random sampling as follows:

i) Simple random sampling with replacement $(srswr)$.

ii) Simple random sampling without replacement $(srswor)$.

Simple random sampling with replacement $(srswr)$

In sampling with replacement a unit is selected from the population consisting of N units, its content noted and then returned to the population before the next draw is made, and the process is repeated n times to give a sample of n units. In this method, at each draw, each of the N units of the population gets the same probability $\frac{1}{N}$ of being selected. Here the same unit of the population may occur more than once in the sample (order in which the sample units are obtained is regarded). There are N^n samples, and each has an equal probability $\frac{1}{N^n}$ of being selected.

Simple random sampling without replacement $(srswor)$

Suppose the population consist of N units, then, in simple random sampling without replacement a unit is selected, its content noted and the unit is not returned to the population before next draw is made. The process is repeated n times to give a sample of n units. In this method at the $r-$th drawing, each of the $N-r+1$ units of the population gets the same probability $\frac{1}{N-r+1}$ of being included in the sample. Here any unit of the population cannot occur more than once in the sample (order is ignored). There are ${}^{N}C_n$ possible samples, and each such sample has an equal probability $\frac{1}{{}^{N}C_n}$ of being selected.

Example: For a population of size $N=5$ with values 1, 3, 6, 8 and 9 make list of all possible samples of size $n=3$ by $srswor$.

Solution:

By the sampling wor, the number of possible samples will be ${}^{N}C_n = {}^{5}C_3 = 10$, which are as follows:

(1, 3, 6), (1, 3, 8), (1, 3, 9), (1, 6, 8), (1, 6, 9), (1, 8, 9), (3, 6, 8), (3, 6, 9), (3, 8, 9), (6, 8, 9).

Theory of simple random sampling with replacement

N, population size.

n, sample size.

Y_i, value of the $i-$th unit of the population.

y_i, value of the $i-$th unit of the sample.

$Y = \sum_{i=1}^{N} Y_i$, population total.

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$, population mean.

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, sample mean.

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}^2$, population variance.

$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right)$, population mean square.

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)$, sample mean square.

Theorem: In *srswr*, the sample mean $\bar{y}$ is an unbiased estimate of the population mean $\bar{Y}$ i.e. $E(\bar{y}) = \bar{Y}$ and its variance $V(\bar{y}) = \frac{N-1}{nN}S^2 = \frac{\sigma^2}{n}$.

Remarks:

i) The standard error (SE) of $\bar{y}$ is $SE(\bar{y}) = \sqrt{V(\bar{y})} = \frac{\sigma}{\sqrt{n}} = S\sqrt{\frac{N-1}{nN}}$.

Theorem: In *srswr*, sample mean square s^2 is an unbiased estimate of the population variance σ^2 i.e. $E(s^2) = \sigma^2$.

Example: In a population with $N = 5$, the values of Y_i are 8, 3, 11, 4 and 7.

a) Calculate population mean $\bar{Y}$, variance σ^2 and mean sum square S^2.

b) Enumerate all possible samples of size 2 by the replacement method and verify that

   i) Sample mean $\bar{y}$ is unbiased estimate of population mean $\bar{Y}$ i.e. $E(\bar{y}) = \bar{Y}$.

   ii) $N\bar{y}$ is unbiased estimate of population total Y i.e. $E(N\bar{y}) = Y$.

   iii) $V(\bar{y}) = \frac{(N-1)S^2}{nN} = \frac{\sigma^2}{n}$, and

   iv) $E(s^2) = \left(\frac{N-1}{N}\right)S^2 = \sigma^2$.

Solution:

a) We know that

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = 6.6$, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}^2 = 8.24$, and $S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right) = 10.3$

b) Form a table for calculation as below:

Samples	$\bar{y}_i$	$\bar{y}_i^2$	$N\,\bar{y}_i$	s_i^2	Samples	$\bar{y}_i$	$\bar{y}_i^2$	$N\,\bar{y}_i$	s_i^2
(8, 8)	8.0	64.00	40.0	0.0	(11, 4)	7.5	56.25	37.5	24.5
(8, 3)	5.5	30.25	27.5	12.5	(11, 7)	9.0	81.00	45.0	8.0
(8, 11)	9.5	90.25	47.5	4.5	(4, 8)	6.0	36.00	30.0	8.0
(8, 4)	6.0	36.00	30.0	8.0	(4, 3)	3.5	12.25	17.5	0.5
(8, 7)	7.5	56.25	37.5	0.5	(4, 11)	7.5	56.25	37.5	24.5
(3, 8)	5.5	30.25	27.5	12.5	(4, 4)	4.0	16.00	20.0	0.0
(3, 3)	3.0	9.00	15.0	0.0	(4, 7)	5.5	30.25	27.5	4.5
(3, 11)	7.0	49.00	35.0	32.0	(7, 8)	7.5	56.25	37.5	0.5
(3, 4)	3.5	12.25	17.5	0.5	(7, 3)	5.0	25.00	25.0	8.0
(3, 7)	5.0	25.00	25.0	8.0	(7, 11)	9.0	81.00	45.0	8.0
(11, 8)	9.5	90.25	47.5	4.5	(7, 4)	5.5	30.25	27.5	4.5
(11, 3)	7.0	49.00	35.0	32.0	(7, 7)	7.0	49.00	35.0	0.0
(11, 11)	11.0	121.00	55.0	0.0					

i) $E(\bar{y}) = \frac{1}{n'}\sum_{i=1}^{n'} \bar{y}_i = \frac{1}{25} \times 165 = 6.6 = \bar{Y}$, where n' is the number of sample.

ii) $E(N\,\bar{y}) = \frac{1}{n'}\sum_{i=1}^{n'} N\,\bar{y}_i = 33$ or $E(N\,\bar{y}) = N\,E(\bar{y}) = 33$.

iii) $V(\bar{y}) = \frac{1}{n'}\sum_{i=1}^{n'} \bar{y}_i^2 - \bar{Y}^2 = 4.12$.

Now,

$$\frac{(N-1)S^2}{nN} = 4.12, \text{ and } \frac{\sigma^2}{n} = 4.12, \text{ therefore,}$$

$$V(\bar{y}) = \frac{(n-1)S^2}{nN} = \frac{\sigma^2}{n} = 4.12.$$

iv) $E(s^2) = \frac{1}{n'}\sum_{i=1}^{n'} s_i^2 = \frac{1}{25} \times 206 = 8.24$ (1a)

and $\frac{(N-1)S^2}{N} = 8.24$ (2a)

In view of equation (1a) and (2a), we get

$$E(s^2) = \frac{(N-1)S^2}{N} = \sigma^2 = 8.24.$$

Theory of simple random sampling without replacement

Theorem: In *srswor*, sample mean $\bar{y}$ is an unbiased estimate of the population mean $\bar{Y}$ i.e. $E(\bar{y}) = \bar{Y}$ and its variance is $V(\bar{y}) = \left(\frac{N-n}{nN}\right)S^2$.

.

Remarks:

i) The standard error of $\bar{y}$ is $SE(\bar{y}) = S\sqrt{\frac{N-n}{nN}} = S\sqrt{\frac{1-f}{n}} = S\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)}$.

Theorem: In *srswor*, sample mean square s^2 is an unbiased estimate of the population mean square S^2 i.e. $E(s^2) = S^2$.

Example: A random sample of $n = 2$ households was drawn from a small colony of $N = 5$ households having monthly income (in rupees) as follows:

Households:	1	2	3	4	5
Income (in thousand rupees):	8	6.5	7.5	7	6

a) Calculate population mean $\bar{Y}$, variance σ^2 and mean sum square S^2.

b) Enumerate all possible samples of size $n = 2$ by the without replacement method and verify that

i) Sample mean $\bar{y}$ is unbiased estimate of population mean $\bar{Y}$ i.e. $E(\bar{y}) = \bar{Y}$.

ii) $N\bar{y}$ is unbiased estimate of population total Y i.e. $E(N\bar{y}) = Y$.

iii) $V(\bar{y}) = \frac{(N-n)S^2}{nN}$, and

iv) $E(s^2) = S^2$.

Solution:

a) We know that

$$\bar{Y} = \frac{1}{N}\sum_{i-1}^{N} Y_i = 7, \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}^2 = 0.5, \text{ and } S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right) = 0.625.$$

b) Form a table for calculation as below:

Samples	$\bar{y}_i$	$\bar{y}_i^2$	$N\bar{y}_i$	s_i^2	Samples	$\bar{y}_i$	$\bar{y}_i^2$	$N\bar{y}_i$	s_i^2
(8, 6.5)	7.25	52.563	36.25	1.125	(8, 7.5)	7.75	60.063	38.75	0.125
(8, 7)	7.50	56.250	37.50	0.500	(8, 6)	7.00	49.000	35.00	2.000
(6.5, 7.5)	7.00	49.000	35.00	0.500	(6.5, 7)	6.75	45.563	33.75	0.125
(6.5, 6)	6.25	39.063	31.25	0.125	(7.5, 7)	7.25	52.563	36.25	0.125
(7.5, 6)	6.75	45.563	33.75	1.125	(7, 6)	6.50	42.250	32.50	0.500

i) $E(\bar{y}) = \frac{1}{n'}\sum_{i=1}^{n'} \bar{y}_i = 7 = \bar{Y}$, where n' is the number of sample.

ii) $E(N\bar{y}) = \frac{1}{n'}\sum_{i=1}^{n'} N\bar{y}_i = 35$, or $E(N\bar{y}) = N\,E(\bar{y}) = 35$.

iii) $V(\bar{y}) = \frac{1}{n'}\sum_{i=1}^{n'} (\bar{y}_i - \bar{Y})^2 = \frac{1}{n'}\sum_{i=1}^{n'} \bar{y}_i^2 - \bar{Y}^2 = 0.1875$, and $\frac{(N-n)S^2}{nN} = 0.1875$.

Therefore,

$$V(\bar{y}) = \frac{(N-n)S^2}{nN} = 0.1875 .$$

iv) $E(s^2) = \frac{1}{n'}\sum_{i=1}^{n'} s_i^2 = 0.625 = S^2$.

Property: $V(\bar{y})$ under *srswor* is less than the $V(\bar{y})$ under *srswr* .

Proof:

Under *srswor* $V(\bar{y}) = \frac{N-n}{nN}S^2$ (1.7)

and under *srswr* $V(\bar{y}) = \frac{\sigma^2}{n} = \frac{N-1}{nN}S^2$ (1.8)

Comparing (1.7) and (1.8), we note that $(N-1) > (N-n)$, which is always the case $\frac{N-1}{nN}S^2 > \frac{N-n}{nN}S^2$.

Example: In a population $N = 5$, the values are 2, 4, 6, 8 and 10, then for a *srs* size $n = 3$, show that $V(\bar{y})_{srswor} < V(\bar{y})_{srswr}$.

Solution: We know that

$$V(\bar{y})_{srswor} = \frac{N-n}{nN}S^2, \text{ and } \quad V(\bar{y})_{srswr} = \frac{N-1}{nN}S^2,$$

where, $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = 10$ and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = 6$.

Thus,

$$V(\bar{y})_{srswor} = \frac{4}{3},\ V(\bar{y})_{srswr} = \frac{8}{3}, \text{ and therefore,}$$

$$V(\bar{y})_{srswor} < V(\bar{y})_{srswr} .$$

SRATIFIED RANDOM SAMPLING

The precision of an estimator of the population parameters (mean or total etc.) depends on the size of the sample and the variability or heterogeneity among the units of the population. If the population is very heterogeneous and considerations of cost limit the size of the sample, it may be found impossible to get a sufficiently precise estimate by taking a simple random sample from the entire population. For this, one possible way to estimate the population mean or total with greater precision is to divide the population in several groups (sub-population or classes, these sub-populations are non-overlapping) each of which is more homogenous than the entire population and draw a random sample of predetermined size from each one of the groups. The groups, into which the population is divided, are called **strata** or each group is called **stratum** and the whole procedure of dividing the population into the strata and then drawing a random sample from each one of the strata is called **stratified random sampling**. For example, to estimate the average income per household, it may be appropriate to group the households into two or more groups (strata) according to the rent paid by the households. The households in any stratum so form are likely to be more homogeneous with respect to income as compared to the whole population. Thus, the estimated income per household based on a stratified sample is likely to be more precise than that based on a simple random sample of the same size drawn from the whole population.

Principal reasons for stratification

- To gain in precision, divide a heterogeneous population into strata in such a way that each stratum is internally homogeneous.
- To accommodate administrative convenience (cost consideration), fieldwork is organized by strata, which usually results in saving in cost and effort.
- To obtain separate estimates for strata.
- We can accommodate different sampling plan in different strata.
- We can have data of known precision for certain subdivisions treating each subdivision as a population in its own right.

Notations: Let the population, consisting of N units is first divided into k strata (sub-populations) of size $N_1, N_2, \cdots, N_k$. These sub-populations are non-overlapping such that $N_1 + N_2 + \cdots + N_k = N$. A sample is drawn (by the method of srs) from each stratum (group or sub-population) independently, the sample size within the $i-$th stratum being n_i, $(i = 1, 2, \cdots, k)$ such that $n_1 + n_2 + \cdots + n_k = n$. The following symbols refer to stratum i.

N_i, total number of units.

n_i, number of units in sample.

$f_i = \dfrac{n_i}{N_i}$, sampling fraction in the stratum.

$W_i = \dfrac{N_i}{N}$, stratum weight.

y_{ij}, value of the characteristic under study for the $j-$th unit in the $i-$th stratum, $j = 1, 2, \cdots, N_i$.

$\bar{Y}_i = \dfrac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$, mean based on N_i units (stratum mean).

$\bar{y}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, mean based on n_i units (sample mean).

$\sigma_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}(y_{ij} - \bar{Y}_i)^2$, variance based on N_i units (stratum variance).

$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(y_{ij} - \bar{Y}_i)^2$, mean square based on N_i units (stratum mean square).

$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$, sample mean square based on n_i units.

$Y = \sum_{i=1}^{k}\sum_{j=1}^{N_i} y_{ij} = \sum_{i=1}^{k} N_i \bar{Y}_i$, population total.

$\bar{Y} = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{Y}_i = \sum_{i=1}^{k} W_i \bar{Y}_i$, over all population mean.

Theorem: For stratified random sampling, *wor* , if in every stratum the sample estimate $\bar{y}_i$ is an unbiased of $\bar{Y}_i$, and samples are drawn independently in different strata, then $\bar{y}_{st} = \sum_{i=1}^{k} W_i \bar{y}_i$ is an unbiased estimate of the over all population mean $\bar{Y}$ and its variance is $V(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2$.

Exercise: In a population with $N = 6$ and $k = 2$ the values of y_{ij} are 0, 1, 2 in stratum 1 and 4, 6, 11 in stratum 2. A sample with $n = 4$ is to be taken.

i) Show that the optimum n_i under Neyman allocation are $n_1 = 1$ and $n_2 = 3$.

ii) Compute the estimate $\bar{y}_{st}$ for every possible sample under optimum allocation and proportion allocation. Show that the estimates are unbiased. Hence find $V(\bar{y}_{st})$ directly under optimum and proportion allocation and verify that $V(\bar{y}_{st})$ under optimum agrees with the formula $V(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} W_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_i S_i^2 = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2$ and $V(\bar{y}_{st})$ under proportion agrees with the formula $V(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i S_i^2$.

Solution: Given $N = 6$, $n = 4$, $k = 2$, and $N_1 = N_2 = 3$, also $\bar{Y}_1 = 1$, and $\bar{Y}_2 = 7$.

i) Under Neyman allocation,

$n_i = n\frac{N_i S_i}{\sum_{i=1}^{k} N_i S_i}$, where $S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(y_{ij} - \bar{Y}_i)^2$, so that,

$S_1^2 = \frac{1}{N_1 - 1}\sum_{j=1}^{3}(y_{1j} - \bar{Y}_1)^2 = 1$, and $S_2^2 = \frac{1}{N_2 - 1}\sum_{j=1}^{3}(y_{2j} - \bar{Y}_2)^2 = 13$.

Therefore,

$$n_1 = n\frac{N_1 S_1}{\sum_i N_i S_i} \cong 1, \text{ and } \quad n_2 = n\frac{N_{21} S_2}{\sum_i N_i S_i} \cong 3.$$

ii) Possible samples under optimum allocation will be ${}^3C_1 \times {}^3C_3 = 3$, since $n_1 = 1$, $n_2 = 3$ and $N_1 = 3$, $N_2 = 3$

Samples		Means		
I	II	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_{st}$
0	(4, 6, 11)	0	7	3.5
1	(4, 6, 11)	1	7	4.0
2	(4, 6, 11)	2	7	4.5

$E(\bar{y}_{st}) = (3.5 + 4.0 + 4.5)/3 = 4 = \bar{Y}$, thus $\bar{y}_{st}$ is unbiased estimate of $\bar{Y}$ under optimum allocation.

$$V(\bar{y}_{st}) = [(3.5-4)^2 + (4.0-4)^2 + (4.5-4)^2]/3 = 0.1667.$$

$$V(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)W_i^2 S_i^2 = \left(\frac{1}{n_1} - \frac{1}{N_1}\right)W_1^2 S_1^2 + \left(\frac{1}{n_2} - \frac{1}{N_2}\right)W_2^2 S_2^2 = 0.1667.$$

$$V(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} W_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_i S_i^2 = 0.1667 .$$

Possible samples under proportional allocation will be ${}^3C_2 \times {}^3C_2 = 9$, since $n_i = nW_i$, so that $n_1 = 2$, $n_2 = 2$.

Samples		Means		
I	II	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_{st}$
(0, 1)	(4, 6)	0.5	5.0	2.75
(0, 1)	(4, 11)	0.5	7.5	4.00
(0, 1)	(6, 11)	0.5	8.5	4.50
(0, 2)	(4, 6)	1.0	5.0	3.00
(0, 2)	(4, 11)	1.0	7.5	4.25
(0, 2)	(6, 11)	1.0	8.5	4.75
(1, 2)	(4, 6)	1.5	5.0	3.25
(1, 2)	(4, 11)	1.5	7.5	4.50
(1, 2)	(6, 11)	1.5	8.5	5.00

$$E(\bar{y}_{st}) = \frac{1}{9}(2.75 + 4.00 + 4.50 + 3.00 + 4.25 + 4.75 + 3.25 + 4.50 + 5.00) = 4 = \bar{Y}$$

Therefore, $\bar{y}_{st}$ is unbiased estimate of $\bar{Y}$ under proportion allocation.

$$V(\bar{y}_{st}) = \frac{1}{9}[(2.75-4)^2 + (4.00-4)^2 + \cdots + (5.00-4)^2] = 0.583.$$

By formula

$$V(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i S_i^2 = 0.583.$$

SYSTEMATIC RANDOM SAMPLING

A sampling technique in which only the first unit is selected with the help of random numbers and the rest get selected automatically according to some pre-determined pattern (in regular spacing pattern) is known as **systematic random sampling**. Suppose N units of the population are numbered from 1 to N in some order. Let $N = nk$, where n is the sample size, and $k = \frac{N}{n}$, being an integer and usually called the **sampling interval**. Draw a random number less than or equal to k, say i, and select the unit with the corresponding serial number and every $k-$th unit in the population thereafter. The resultant sample will contain the n units with serial numbers $i, i+k, i+2k, \cdots, i+(n-1)k$ is called every $k-$th systematic sample and such a procedure termed **linear systematic sampling**.

Example: There are 50 houses on a street. If a sample of size 5 is to be chosen, then $k = \frac{N}{n} = 10$, we select randomly one house out of the first ten, suppose we select 3rd and then take every 10th house after selected one, i.e. as3rd, 13th, 23rd, 33rd, 43rd. And the possible sample in (linear) systematic sampling will be:

Sample number	Sample units				
1	1	11	21	31	41
2	2	12	22	32	42
3	3	13	23	33	43
⋮	⋮	⋮	⋮	⋮	⋮
10	10	20	30	40	50

Linear systematic sampling suffers from the limitation that it cannot be used when the sampling interval $k = \frac{N}{n}$ is not an integer. The procedure to be followed is that of **circular systematic sampling** [proposed by Lahiri]. In this method, select first item randomly out of N units and then take every $k-$th unit thereafter (where k is the nearest integer to N/n) in a cyclical manner until n sampling units are obtained.

Example: Consider a population of size $N = 33$. If a sample of size $n = 5$ is to be drawn. Here the sampling interval $\frac{N}{n} = 6.6$, is a fractional number, and we have to go in for circular systematic sampling. Since the integer k nearest to 6.6 is 7, we select randomly one item out of $N = 33$, suppose 10th item be selected and then take every 7th item from that one, i.e. as 10th, 17th, 24th, 31st, 6th.

Systematic random sampling is extensively used on account of its low cost and simplicity in the selection of the sample. Apart from its simplicity, this procedure in many situations provides estimates more efficient than simple random sampling. For example, the products coming out continuously from a manufacturing process may be sampled systematically by selecting products manufacturing at a fixed interval of time. Again, in sampling from plants growing in rows, one may devide the whole area into a number of equal rectangular blocks, choose a plant at random from the first block and draw plants exactly from similar spots from other blocks. (It has been found very useful for sampling material continuously distributed over time or in space).