Unit-II MEASURES OF CENTRAL TENDENCY AND DISPERSION

Ungrouped Data (Raw Data): The information collected systematically regarding a population or a sample survey is called an ungrouped data. It is also called raw data.

Grouped Data (Classified Data): When a frequency distribution is obtained by dividing an ungrouped data in a number of strata according to the value of variate under study, such information is called grouped data or classified data.

Measures of central tendency:

There are several measures of central tendency. Out of these, the following 3 are used more often

1. Mean, 2. Median and 3. Mode

Ungrouped Data

Mean or arithmetic mean of ungrouped data:

Let $x_1, x_2, x_3, \dots, x_n$ be n observations then mean is obtained by dividing the sum of n observations by n. It is denoted by

$$\bar{x} = \frac{\sum x_i}{n}$$

Example:

Find the mean of 4,6,8,6,7,8

Solution:

$$\bar{x} = \mu = \frac{\sum x_i}{n}$$

$$= \frac{(4 + 6 + 8 + 6 + 7 + 8)}{6}$$

$$= \frac{39}{6}$$

$$\bar{x} = 6.5$$

Geometric mean:

The average of a set of products, the calculation of which is commonly used to determine the performance results of an investment or portfolio.

The geometric mean of a data set $\{x_1, x_2, x_3, \dots, x_n\}$ is given by:

$$GM = \{x_1\, x_2\, x_3 \dots x_n\}^{\frac{1}{n}}\ or$$

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

$$GM = 10^g \ \ where\ g = \ \frac{1}{n}\sum \log x_i$$

Example:

Find the geometric mean of 4, 6, 8, 6, 7, and 8.

Solution: Here $n = 5$

$$GM = 10^g \ \ where\ g = \ \frac{1}{n}\sum \log x_i$$

$$g = \frac{1}{5}\,(\log 4 + \log 6 + \log 8 + \log 6 + \log 7 + \log\ 8)$$

$$g = \frac{1}{5}\,(0.6021 + 0.7782 + 0.9031 + 0.7782 + 0.8451 + 0.9031)$$

$$g = \frac{1}{5}\,(4.8098) = 0.962$$

$$GM = 10^{0.962} = 9.1622$$

Harmonic mean:

The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals.

The harmonic mean *H* of the positive real numbers $x_1, x_2, x_3, \ldots, x_n > 0$ is defined to be

$$H = \frac{n}{\sum \frac{1}{x_i}}$$

Example:

A gas-powered pump can drain a pool in 4 hours and a battery-powered pump can drain the same pool in 6 hours, then how long it will take both pumps to drain the pool together.

Solution: It is one-half of the harmonic mean of 6 and 4.

$$H = \frac{n}{\sum \frac{1}{x_i}} = \frac{2}{\frac{1}{4} + \frac{1}{6}} = \frac{2 \times 6 \times 4}{6 + 4}$$

$$H = \frac{48}{10} = 4.8$$

Therefore it will take both pumps $\frac{4.8}{2} = 2.4$ hours, to drain the pool together.

Example:

Find the harmonic mean of 4, 6, 8, 6, 7, and 8.

Solution*:* Here $n = 5$

$$H = \frac{n}{\sum \frac{1}{x_i}} = \frac{5}{\frac{1}{4} + \frac{1}{6} + \frac{1}{8} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}} = \frac{5}{0.25 + 0.167 + 0.125 + 0.167 + 0.143 + 0.125}$$

$$H = \frac{5}{0.977} = 5.118$$

Relation between arithmetic, geometric and harmonic means:

i) $\text{Harmonic mean} \leq \text{Geometric mean} \leq \text{Arithmetic mean}$

ii) $\text{Geometric mean} = \sqrt{\text{Arithmetic mean} \times \text{Harmonic mean}}$

Weighted mean*:*

The weighted mean of a set of numbers $x_1, x_2, x_3, \ldots, x_n$ with corresponding weights $w_1, w_2, w_3, \ldots, w_n$ is computed from the following formula:

$$Weighted\ mean\ WM = \frac{\sum w_i x_i}{\sum w_i}$$

Example:

During a one hour period on a hot Saturday afternoon cabana boy Chris served fifty drinks. He sold five drinks for $0.50, fifteen for $0.75, fifteen for $0.90, and fifteen for $1.10. Compute the weighted mean of the price of the drinks.

Solution:

$$Weighted\ mean\ WM = \frac{\sum w_i x_i}{\sum w_i} = \frac{5 \times 0.50 + 15 \times 0.75 + 15\ \times 0.90 + 15\ \times 1.10}{5 + 15 + 15 + 15}$$

$$WM\ = \frac{44.34}{50}$$

$$WM\ =\ 0.89$$

Median of ungrouped data:

If the observations of an ungrouped data are arranged in increasing or decreasing order of their magnitude, a value which divides these ordered observations into two equal parts is called the median of the data. It is denoted by M.

If the number of observations (n) is an **odd** integer, then

$$M\ =\ Value\ of\ \frac{(n+1)^{th}}{2}\ observation$$

If the number of observations (n) is an **even** integer, then

$$M\ = \frac{(Value\ of\ \frac{n^{th}}{2}\ observation\ +\ Value\ of\ \frac{(n+1)^{th}}{2}\ observation)}{2}$$

Example:

Find the median of the following observations 4,6,8,6,7,8,8

Solution: Observations in the ascending order are :
4, 6, 6, 7, 8, 8, 8
Here, $n\ =\ 7$ is odd.
Median :

$$M\ =\ Value\ of\ \frac{(n+1)^{th}}{2}\ observation$$

$$M\ =\ Value\ of\ \frac{(7+1)}{2}^{th}\ observation$$

$$M\ =\ Value\ of\ \ 4^{th}\ observation$$

$$M\ =\ 7$$

Note: The **median** divides the data into a lower half and an upper half.

Mode of ungrouped data:

An observation occurring most frequently in the data is called mode of the data. It is denoted by Z.

Example:

Find the median of the following observations
4,6,8,6,7,8,8

Solution: In the given data, the observation 8 occurs maximum number of times (3)

$$Mode(Z)\ =\ 8$$

Unit-II MEASURES OF CENTRAL TENDENCY AND DISPERSION

Relation between Mean, Median and Mode:

$$Mean - Mode = 3(Mean - Median)$$

Range of ungrouped data:

The range of a set of data is the difference between the highest and lowest values in the set.

Example:

Cheryl took 7 math tests in one marking period. What is the range of her test scores?

89, 73, 84, 91, 87, 77, 94

Solution: Ordering the test scores from least to greatest, we get: 73, 77, 84, 87, 89, 91, and 94

Highest - lowest = 94 - 73 = 21

Therefore the range of these test scores is 21 points.

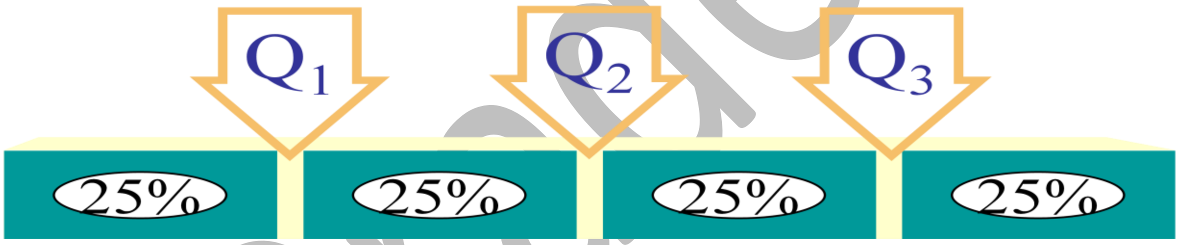
Quartile of ungrouped data:

Quartile - measures of central tendency that divide a group of data into four subgroups

Q_1: 25% of the data set is below the first quartile

Q_2: 50% of the data set is below the second quartile

Q_3: 75% of the data set is below the third quartile



The lower quartile or first quartile is the middle value of the lower half.

The upper quartile or third quartile is the middle value of the upper half.

Example:

Find the median, lower quartile and upper quartile of the following numbers.

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

Solution:

First, arrange the data in ascending order:

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

Median (middle value) = 22

Lower quartile (middle value of the lower half) = 12

Upper quartile (middle value of the upper half) = 36

If there is an even number of data items, then we need to get the average of the middle numbers.

Interquartile range:

Interquartile range = Upper quartile – lower quartile

Unit-II MEASURES OF CENTRAL TENDENCY AND DISPERSION

Example:

Find the median, lower quartile, upper quartile, interquartile range and range of the following numbers.

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

Solution:

First, arrange the data in ascending order:

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

Lower quartile or first quartile = $\frac{12+14}{2} = 13$

Median or second quartile = $\frac{22+25}{2} = 23.5$

Upper quartile or third quartile = $\frac{36+42}{2} = 39$

Interquartile range = Upper quartile – lower quartile

= 39 – 13 = 26

Range = largest value – smallest value

= 65 – 5 = 60

***Variance of an ungrouped data*:**

Variance is the average of the squared deviations from the arithmetic mean.

The variance of a set of values, which we denote by σ^2, is defined as

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

Where $\bar{x}$ is the mean, n is the number of data values, and x_i stands for data value in i^th^ position.

An alternative, yet equivalent formula, which is often easier to use is

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

Example:

Find the variance of 6, 7, 10, 11, 11, 13, 16, 18, and 25.

Solution: Here $n = 9$

Firstly we find the mean,

$$\bar{x} = \frac{\sum x_i}{n} = \frac{6+7+\cdots+25}{9} = \frac{117}{9} = 13$$

Method 1:

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

It is helpful to show the calculation in a table:

x_i	6	7	10	11	11	13	16	18	25	Total
$x_i - \bar{x}$	-7	-6	-3	-2	-2	0	3	5	12	
$(x_i - \bar{x})^2$	49	36	9	4	4	0	9	25	144	280

$$\sigma^2 = \frac{280}{9} = 31.11$$

Method 2:

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

x_i	6	7	10	11	11	13	16	18	25	Total
${x_i}^2$	36	49	100	121	121	169	256	324	625	1801

$$= \frac{\sum(x_i)^2}{n} - \bar{x}^2 = \frac{1801}{9} - 13^2$$

$$= 200.11 - 169 = 31.11$$

Standard Deviation (σ)

Since the variance is measured in terms of x_i^2, we often wish to use the standard deviation where

$$\sigma = \sqrt{Variance}$$

The standard deviation, unlike the variance, will be measured in the same units as the original data.

$$\sigma = \sqrt{31.11} = 5.58$$

Coefficient of Variation:

Coefficient of Variation (CV) is the ratio of the standard deviation to the mean, expressed as a percentage.

$$CV = \frac{\sigma}{\mu} \times 100$$

Example:

Find the Coefficient of Variation of 6, 7, 10, 11, 11, 13, 16, 18, and 25.

Solution: Here $n = 9$

Firstly we find the mean,

$$\bar{x} = \mu = \frac{\sum x_i}{n} = \frac{6 + 7 + \cdots + 25}{9} = \frac{117}{9} = 13$$

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

It is helpful to show the calculation in a table:

x_i	6	7	10	11	11	13	16	18	25	Total
$x_i - \bar{x}$	-7	-6	-3	-2	-2	0	3	5	12	
$(x_i - \bar{x})^2$	49	36	9	4	4	0	9	25	144	280

$$\sigma^2 = \frac{280}{9} = 31.11$$

$$CV = \frac{\sigma}{\mu} \times 100 = \frac{\sqrt{31.11}}{13} \times 100$$

$$CV = \frac{5.58}{13} \times 100 = 42.92$$

Mean Absolute deviation:

Mean Absolute Deviation is the average of the absolute deviations from the mean.

$$M.A.D = \frac{\sum |x_i - \bar{x}|}{n}$$

Example:

Find the Mean Absolute Deviation of 6, 7, 10, 11, 11, 13, 16, 18, and 25.

Solution: Here $n = 9$

Firstly we find the mean,

$$\bar{x} = \mu = \frac{\sum x_i}{n} = \frac{6 + 7 + \cdots + 25}{9} = \frac{117}{9} = 13$$

It is helpful to show the calculation in a table:

x_i	6	7	10	11	11	13	16	18	25	Total
$\lvert x_i - \bar{x}\rvert$	7	6	3	2	2	0	3	5	12	40

$$M.A.D = \frac{\sum |x_i - \bar{x}|}{n} = \frac{40}{9} = 4.44$$

Grouped Data:

Mean:

For the grouped data the class interval and frequency is given we have to find the midpoint of class interval and find the mean using the formula

$$\text{Mean } \bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

Where x_i is the midpoint of the ith class interval,

f_i is the frequency of the ith class interval.

Median:

The median for the grouped data is also defined as the value corresponding to the $\frac{n+1}{2}^{th}$ observation, and is calculated from the formula:

$$\text{Median} = L_m + \frac{\left(\frac{n}{2} - f_c\right)}{f_m} \times W_m$$

Where L_m is the lower limit of the median class interval i.e. the interval which contains $\frac{n}{2}^{th}$ observation,

f_m is the frequency of the median class interval,

f_c is the cumulative frequency up to the interval just before the median class interval,

W_m is the width of the median class interval, and n is the number of total observations i.e, $n = \sum f_i$.

Mode:

In a grouped data, the mode is calculated by the following formula

$$\text{Mode} = L_m + \frac{f_m - f_0}{2f_m - f_0 - f_2} \times W_m$$

Where L_m is the lower limit of the modal class interval i.e. the class interval having highest frequency,

f_m is the frequency of the modal class interval,

f_0 is the frequency of to the interval just before the modal class interval,

f_2 is the frequency of to the interval just after the modal class interval

W_m is the width of the modal class interval

First Quartile:

$$\text{First Quartile } Q_1 = L_{Q_1} + \frac{\left(\frac{n}{4} - f_c\right)}{f_{Q_1}} \times W_{Q_1}$$

Where L_{Q_1} is the lower limit of the first quartile class interval i.e. the interval which contains $\frac{n}{4}^{th}$ observation,
f_{Q_1} is the frequency of the first quartile class interval,
f_c is the cumulative frequency up to the interval just before the first quartile class interval,
W_{Q_1} is the width of the first quartile class interval, and n is the number of total observations i.e, $n = \sum f_i$.

Third Quartile:

$$\text{Third Quartile } Q_3 = L_{Q_3} + \frac{\left(\frac{3n}{4} - f_c\right)}{f_{Q_3}} \times W_{Q_3}$$

Where L_{Q_3} is the lower limit of the third quartile class interval i.e. the interval which contains $\frac{3n}{4}^{th}$ observation,
f_{Q_3} is the frequency of the third quartile class interval,
f_c is the cumulative frequency up to the interval just before the first quartile class interval,
W_{Q_3} is the width of the third quartile class interval, and n is the number of total observations i.e, $n = \sum f_i$.

Quartile Deviation or Semi Inter-Quartile Range:

$$Quartile\ Deviation = \frac{Q_3 - Q_1}{2}$$

Percentiles:

Percentiles which spilt the data into several parts, expressed in percentages. A Percentile also known as centile, divides the data in such a way that “given percent of the observations are less than it”.

$$P_m = L_{P_m} + \frac{\left(\frac{mn}{100} - f_c\right)}{f_{P_m}} \times W_{P_m}$$

WhereL_{P_m} is the lower limit of the percentile class interval i.e. the interval which contains $\frac{mn}{100}^{th}$ observation,
f_{P_m} is the frequency of the percentile class interval,
f_c is the cumulative frequency up to the interval just before the first quartile class interval,
W_{P_m} is the width of the percentile class interval, and n is the number of total observations i.e, $n = \sum f_i$.

Deciles:

The deciles divide the data into ten parts- first decile (10%), second (20%) and so on**.**

$$\text{Decile } D_m = L_{D_m} + \frac{\left(\frac{mn}{10} - f_c\right)}{f_{D_m}} \times W_{D_m}$$

WhereL_{D_m} is the lower limit of the percentile class interval i.e. the interval which contains $\frac{mn}{10}^{th}$ observation,
f_{D_m} is the frequency of the percentile class interval,
f_c is the cumulative frequency up to the interval just before the first quartile class interval,
W_{D_m} is the width of the percentile class interval, and n is the number of total observations i.e, $n = \sum f_i$

Mean Absolute Deviation:

$$M.A.D = \frac{\sum(f_i|x_i - \bar{x}|)}{\sum f_i}$$

Where x_i is the midpoint of the ith class interval,
f_i is the frequency of the ith class interval.
$\bar{x}$ is the mean of the given data

Variance:

$$\sigma^2 = \frac{\sum f_i x_i^2 - (\sum f_i)\,\bar{x}^2}{\sum f_i}$$

Where x_i is the midpoint of the ith class interval,
f_i is the frequency of the ith class interval.
$\bar{x}$ is the mean of the given data

Example:

Find Mean, Variance, standard deviation and Coefficient of variation for the following data given below:

Class Interval:	**2000-3000**	**3000-4000**	**4000-5000**	**5000-6000**	**6000-7000**
Frequency:	**2**	**5**	**6**	**4**	**3**

Solution:

Class Interval	**Frequency** f_i	**Midpoint of Class Interval** x_i	x_i^2	$f_i x_i$	$f_i x_i^2$
2000-3000	2	2500	6250000	5000	12500000
3000-4000	5	3500	12250000	17500	61250000
4000-5000	6	4500	20250000	27000	121500000
5000-6000	4	5500	30250000	22000	121000000
6000-7000	3	6500	42250000	19500	126750000
SUMS	**20**			**91000**	**443000000**

$$\text{Mean } \bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{91000}{20} = 4550$$

$$\text{Variance}\ \sigma^2 = \frac{\sum f_i x_i^2 - (\sum f_i)\,\bar{x}^2}{\sum f_i} = \frac{443000000 - 20(4550)^2}{20} = 1447500$$

$$\text{Standard deviation}\ \ \sigma = \sqrt{1447500}\ = 1203.12$$

$$\text{Coefficient of Variation}\ \ \text{C.V} = \frac{\sigma}{\bar{x}} \times 100 = \frac{1203.12}{4550}\ \times 100 = 26.44\ \%$$

Example:

The distribution of Intelligence Quotient (I.Q.) scores measured for 100 students in a test is as follows

I.Q. :	**40-50**	**50-60**	**60-70**	**70-80**	**80-90**	**90-100**
Number of students :	**10**	**20**	**20**	**15**	**15**	**20**

Find Mean, Median, Mode, first Quartile, third Quartile, 60th Percentile, 9th Decile, and Mean deviation.

Solution:

Class Interval	**Frequency** f_i	**Midpoint of Class Interval** x_i	$f_i x_i$	**Cumulative Frequency** f_c	$\lvert x_i - \bar{x}\rvert = \lvert x_i - 71.5\rvert$	$f_i \lvert x_i - \bar{x}\rvert$
40-50	10	45	450	10	26.5	265
50-60	20**	55	1100	30***	16.5	330
60-70	20	65	1300	50*	6.5	130
70-80	15	75	1125	65#	3.5	52.5
80-90	15	85	1275	80****	13.5	202.5
90-100	20	95	1900	100 ##	23.5	470
SUMS	**100**		**7150**			**1450**

To find Mean:

$$\text{Mean}\ \bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{7150}{100} = 71.5$$

To find Median:

$$n = \sum f_i = 100$$

$$\frac{n}{2} = \frac{100}{2} = 50$$

The Median class interval * is 60-70

$(\,Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{n}{2}^{th}\ f_c\ column)$

$$\text{Median} = \ L_m + \frac{\left(\frac{n}{2} - \ f_c\right)}{f_m} \times W_m\ = 60 + \frac{\left(\frac{100}{2} - \ 30\right)}{20} \times (70 - 60)$$

$$= 60 + \frac{(50 - \ 30)}{20} \times 10$$

$$\text{Median} = 70$$

To find Mode:
The Mode class interval ** is 50-60 $(\,Corresponding\ class\ interval\ of\ highest\ value\ in\ f_i\ column)$

$$\text{Mode} = \ L_m + \frac{f_m - f_0}{2f_m - f_0 - f_2} \times W_m$$

$$= \ 50 + \frac{20 - \ 10}{2(20) - \ 10 - \ 20} \times (60 - 50)$$

$$= \ 50 + \frac{10}{10} \times 10$$

$$\text{Mode} = 60.$$

To find First Quartile:

$$n = \sum f_i = 100$$

$$\frac{n}{4} = \frac{100}{4} = 25$$

The First Quartile class interval *** is 50-60
$(\,Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{n}{4}^{th}\ f_c\ column)$

$$\text{First Quartile } Q_1 \ = \ L_{Q_1} + \frac{\left(\frac{n}{4} - \ f_c\right)}{f_{Q_1}} \times \ W_{Q_1} = 50 + \frac{\left(\frac{100}{4} - \ 10\right)}{20} \times (60 - 50)$$

$$= 50 + \frac{(25 - \ 10)}{20} \times 10$$

$$\text{First Quartile } Q_1 \ = 57.5$$

To find Third Quartile:

$$n = \sum f_i = 100$$

$$\frac{3n}{4} = \frac{3(100)}{4} = 75$$

The Third Quartile class interval **** is 80-90
$(\,Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{3n}{4}^{th}\ f_c\ column)$

$$\text{Third Quartile } Q_3 \ = \ L_{Q_3} + \frac{\left(\frac{3n}{4} - \ f_c\right)}{f_{Q_3}} \times \ W_{Q_3} = 80 + \frac{\left(\frac{3(100)}{4} - \ 65\right)}{15} \times (90 - 80)$$

$$= 80 + \frac{(75 - \ 65)}{15} \times 10$$

$$\text{Third Quartile } Q_3 \ = 86.67$$

To find Percentile :

$$P_m = L_{P_m} + \frac{\left(\frac{mn}{100} - f_c\right)}{f_{P_m}} \times W_{P_m}$$

$$\frac{mn}{100} = \frac{60 \times 100}{100} = 60$$

The 60th Percentile class interval # is 70 - 80

$$(\,Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{mn^{th}}{100}\ f_c\ column)$$

$$60^{th}\text{ Percentile } P_{60} = 70 + \frac{(60 - 50)}{15} \times (80 - 70)$$

$$60^{th}\text{ Percentile } P_{60} = 76.67$$

To find Decile:

$$\text{Decile } D_m = L_{D_m} + \frac{\left(\frac{mn}{10} - f_c\right)}{f_{D_m}} \times W_{D_m}$$

$$\frac{mn}{10} = \frac{9 \times 100}{10} = 90$$

The 9th Decile class interval ## is 90 - 100

$$(\,Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{mn^{th}}{10}\ f_c\ column)$$

$$9^{th}\text{ Decile } D_9 = 90 + \frac{(90 - 80)}{20} \times (100 - 90)$$

$$9^{th}\text{ Decile } D_9 = 95$$

To find Mean deviation:

$$\text{Mean deviation} = \frac{\sum f_i|x_i - \bar{x}|}{\sum f_i} = \frac{1450}{100} = 14.5$$

Example:

Calculate the mean, median, mode, 70th percentile and 4th decile from the following data:

Class-interval:	**0-4**	**4-8**	**8-12**	**12-16**	**16-20**	**20-24**	**24-28**
Frequency :	**10**	**12**	**18**	**7**	**5**	**8**	**4**

Solution:

Class Interval	**Frequency** f_i	**Midpoint of Class Interval** x_i	$f_i x_i$	**Cumulative Frequency** f_c
0 - 4	10	2	20	10
4 - 8	12	6	72	22
8 - 12	18**	10	180	40*##
12 - 16	7	14	98	47#
16 - 20	5	18	90	52
20 - 24	8	22	176	60
24 - 28	4	26	104	64
Sums	**64**		**740**	

Unit-II MEASURES OF CENTRAL TENDENCY AND DISPERSION

To find Mean:

$$\text{Mean } \bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{740}{64} = 11.56$$

To find Median:

$$n = \sum f_i = 64$$

$$\frac{n}{2} = \frac{64}{2} = 32$$

The Median class interval * is 8 - 12

$(\, Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{n}{2}^{th}\ f_c\ column)$

$$\text{Median} = L_m + \frac{\left(\frac{n}{2} - f_c\right)}{f_m} \times W_m = 8 + \frac{(32 - 22)}{18} \times (12 - 8)$$

$$= 8 + \frac{10}{18} \times 4$$

$$\text{Median} = 10.22$$

To find Mode:

The Mode class interval ** is 8 - 12 $(\, Corresponding\ class\ interval\ of\ highest\ value\ in\ f_i\ column)$

$$\text{Mode} = L_m + \frac{f_m - f_0}{2f_m - f_0 - f_2} \times W_m$$

$$= 8 + \frac{18 - 12}{2(18) - 12 - 7} \times (12 - 8)$$

$$= 8 + \frac{6}{17} \times 4$$

$$\text{Mode} = 9.41$$

To find Percentile :

$$P_m = L_{P_m} + \frac{\left(\frac{mn}{100} - f_c\right)}{f_{P_m}} \times W_{P_m}$$

$$\frac{mn}{100} = \frac{70 \times 64}{100} = 44.8$$

The 70th Percentile class interval # is 12 - 16

$(\, Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{mn}{100}^{th}\ f_c\ column)$

$$70^{th}\text{ Percentile } P_{70} = 12 + \frac{(44.8 - 40)}{7} \times (16 - 12)$$

$$70^{th}\text{ Percentile } P_{70} = 14.74$$

To find Decile:

$$\text{Decile } D_m = L_{D_m} + \frac{\left(\frac{mn}{10} - f_c\right)}{f_{D_m}} \times W_{D_m}$$

$$\frac{mn}{10} = \frac{4\ \times 64}{10} = 25.6$$

The 4th Decile class interval ## is 8 - 12

$$(\,Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{mn^{th}}{10}\ f_c\ column)$$

$$4^{\text{th}}\text{ Decile } D_4 = \ 8 + \frac{(25.6 - 22)}{18} \times (12 - 8)$$

$$4^{\text{th}}\text{ Decile } D_4 = 8.8$$

Example:

Calculate mean, median, mode, first quartile, third quartile,45th percentile and 9th decile, variance and Coefficient of variation from the following data relating to production of a steel mill on 60 days.

Production (in Tons per day) :	**21-22**	**23-24**	**25-26**	**27-28**	**29-30**
Number of days :	**7**	**13**	**22**	**10**	**8**

Solution: Since the data is not continuous in production we make the data continuous by dividing the width of the interval by 2 and subtracting it from lower value or left value of the interval and adding it in higher value or right value of the interval. Here the width is 1 dividing it by 2 we get 0.5 so subtract 0.5 from lower value or left value of the interval and add 0.5 to higher value or right value of the interval.

Class interval	Frequency f_i	Midpoint of Class Interval x_i	$f_i x_i$	$x_i{}^2$	$f_i x_i{}^2$	Cumulative Frequency f_c
20.5- 22.5	7	21.5	150.5	462.25	3235.75	7
22.5-24.5	13	23.5	305.5	552.25	7179.25	20***
24.5-26.5	22**	25.5	561	650.25	14305.5	42*#
26.5-28.5	10	27.5	275	756.25	7562.5	52****
28.5-30.5	8	29.5	236	870.25	6962	60##
Sum	60	127.5	1528	3291.25	39245	

To find Mean:

$$\text{Mean } \bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1528}{60} = 25.467$$

To find Median:

$$n = \sum f_i = 60$$

$$\frac{n}{2} = \frac{60}{2} = 30$$

The Median class interval * is 24.5 - 26.5

$$(\,Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \tfrac{n^{th}}{2}\ f_c\ column)$$

$$\text{Median } = \ L_m + \frac{\left(\frac{n}{2} - f_c\right)}{f_m} \times W_m = 24.5 + \frac{\left(\frac{60}{2} - 20\right)}{22} \times (26.5 - 24.5)$$

$$= 24.5 + \frac{(30 - 20)}{22} \times 2$$

$$\text{Median} = 25.41$$

To find Mode:

The Mode class interval ** is 24.5 - 26.5

$(\,Corresponding\ class\ interval\ of\ highest\ value\ in\ f_i\ column)$

$$\text{Mode} = L_m + \frac{f_m - f_0}{2f_m - f_0 - f_2} \times W_m$$

$$= 24.5 + \frac{22 - 13}{2(22) - 13 - 10} \times (26.5 - 24.5)$$

$$= 24.5 + \frac{9}{21} \times 2$$

$$\text{Mode} = 25.36$$

To find First Quartile:

$$n = \sum f_i = 100$$

$$\frac{n}{4} = \frac{60}{4} = 15$$

The First Quartile class interval *** is 22.5 - 24.5

$(\,Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{n}{4}^{th}\ f_c\ column)$

$$\text{First Quartile } Q_1 = L_{Q_1} + \frac{\left(\frac{n}{4} - f_c\right)}{f_{Q_1}} \times W_{Q_1} = 22.5 + \frac{\left(\frac{60}{4} - 7\right)}{13} \times (24.5 - 22.5)$$

$$= 22.5 + \frac{(15 - 7)}{13} \times 2$$

$$\text{First Quartile } Q_1 = 23.73$$

To find Third Quartile:

$$n = \sum f_i = 60$$

$$\frac{3n}{4} = \frac{3(60)}{4} = 45$$

The Third Quartile class interval **** is 26.5 - 28.5

$(\,Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{3n}{4}^{th}\ f_c\ column)$

$$\text{Third Quartile } Q_3 = L_{Q_3} + \frac{\left(\frac{3n}{4} - f_c\right)}{f_{Q_3}} \times W_{Q_3} = 26.5 + \frac{\left(\frac{3(60)}{4} - 42\right)}{10} \times (28.5 - 26.5)$$

$$= 26.5 + \frac{(45 - 42)}{10} \times 2$$

$$\text{Third Quartile } Q_3 = 27.1$$

To find Percentile :

$$P_m = L_{P_m} + \frac{\left(\frac{mn}{100} - f_c\right)}{f_{P_m}} \times W_{P_m}$$

$$\frac{mn}{100} = \frac{45 \times 60}{100} = 27$$

The 45th Percentile class interval # is 24.5 - 26.5

$$(\,Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{mn}{100}^{th}\ f_c\ column)$$

$$45^{th}\text{ Percentile } P_{45} = 24.5 + \frac{(27 - 20)}{22} \times (26.5 - 24.5)$$

$$45^{th}\text{ Percentile } P_{45} = 25.14$$

To find Decile:

$$\text{Decile } D_m = L_{D_m} + \frac{\left(\frac{mn}{10} - f_c\right)}{f_{D_m}} \times W_{D_m}$$

$$\frac{mn}{10} = \frac{9 \times 60}{10} = 54$$

The 9th Decile class interval ## is 28.5 – 30.5

$$(\,Corresponding\ class\ interval\ of\ the\ nearest\ greater\ than\ or\ equal\ to\ \frac{mn}{10}^{th}\ f_c\ column)$$

$$9^{th}\text{ Decile } D_9 = 28.5 + \frac{(54 - 52)}{8} \times (30.5 - 28.5)$$

$$9^{th}\text{ Decile } D_9 = 29$$

$$\text{Variance } \sigma^2 = \frac{\sum f_i x_i^2 - (\sum f_i)\,\bar{x}^2}{\sum f_i} = \frac{39245 - 60(25.467)^2}{60} = 5.52$$

$$\text{Standard deviation } \sigma = \sqrt{5.52} = 2.35$$

$$\text{Coefficient of Variation C.V} = \frac{\sigma}{\bar{x}} \times 100 = \frac{2.35}{25.467} \times 100 = 9.23\ \%$$

Estimate:

In statistics, estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample.

Statisticians use sample statistics to estimate population parameters.

For example, sample means are used to estimate population means; sample proportions, to estimate population proportions.

An estimate of a population parameter may be expressed in two ways:

i) Point estimate and ii) Interval estimate.

Point estimate: A point estimate of a population parameter is a single value of a statistic. For example, the sample mean $\bar{x}$ is a point estimate of the population mean μ. Similarly, the sample proportion p is a point estimate of the population proportion P.

It is desirable for a point estimate to be:

(1) **Consistent:** The larger the sample size, the more accurate the estimate.

(2) **Unbiased:** The expectation of the observed values of many samples (“average observation value”) equals the corresponding population parameter.

For example, the sample mean is an unbiased estimator for the population mean.

(3) **Most efficient or best unbiased**: The one possessing the smallest variance (a measure of the amount of dispersion away from the estimate).
In other words, the estimator that varies least from sample to sample. This generally depends on the particular distribution of the population.
For example, the mean is more efficient than the median (middle value) for the normal distribution but not for more "skewed" (asymmetrical) distributions.

Methods to calculate the estimator:

The most commonly used methods to calculate the estimator are
i) **The maximum likelihood method:** Maximum likelihood method uses differential calculus to determine the maximum of the probability function of a number of sample parameters.
ii) **The method of moments:** The moments method equates values of sample moments (functions describing the parameter) to population moments.

Interval estimate: An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, $a < \bar{x} < b$ is an interval estimate of the population mean μ. It indicates that the population mean is greater than a but less than b.

Confidence Intervals:

A confidence interval consists of three parts.
i) Confidence level, ii) Statistic and iii) Margin of error.
The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the *sample statistic* ± *margin of error*.
For example, we might say that we are 95% confident that the true population mean falls within a specified range. This statement is a confidence interval. It means that if we used the same sampling method to select different samples and compute different interval estimates, the true population mean would fall within a range defined by the *sample statistic* ± *margin of error* 95% of the time.
Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

Confidence Level:

The probability part of a confidence interval is called a confidence level. The confidence level describes how strongly we believe that a particular sampling method will produce a confidence interval that includes the true population parameter.
Here is how to interpret a confidence level. Suppose we collected many different samples, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

Margin of Error:

In a confidence interval, the range of values above and below the sample statistic is called the margin of error.
For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.
Note: Many public opinion surveys report interval estimates, but not confidence intervals. They provide the margin of error, but not the confidence level. To clearly interpret survey results you need to know

both! We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%).

Standard error:

The standard error of a statistic is the standard deviation of the sampling distribution of that statistic. Standard errors are important because they reflect how much sampling fluctuation a statistic will show. The inferential statistics involved in the construction of confidence intervals and significance testing are based on standard errors. The standard error of a statistic depends on the sample size. In general, the larger the sample size the smaller the standard error. The standard error of a statistic is usually designated by the Greek letter sigma (σ) with a subscript indicating the statistic.

Standard Error of the Mean:

The standard error of the mean is designated as: σ_M. It is the standard deviation of the sampling distribution of the mean. The formula for the standard error of the mean is:

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the original distribution and n is the sample size (the number of scores each mean is based upon). This formula does not assume a normal distribution. However, many of the uses of the formula do assume a normal distribution. The formula shows that the larger the sample size, the smaller the standard error of the mean. More specifically, the size of the standard error of the mean is inversely proportional to the square root of the sample size.

Standard error of the difference between means:

The standard error of the difference between means is designated as: σ_{M_d} and is given by

$$\sigma_{M_d} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Standard error of the proportion:

The standard error of proportion is designated as: σ_P and is given by

$$\sigma_P = \sqrt{\frac{p_0 q_0}{n}} \text{ where } q_0 = 1 - p_0$$

Standard error of the difference between proportions:

The standard error of the difference between proportions is designated as: σ_{P_d} and is given by

$$\sigma_{P_d} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \text{ where } q_1 = 1 - p_1 \text{ and } q_2 = 1 - p_2$$

Confidence interval:

Confidence interval of mean $(\boldsymbol{\sigma}\ \boldsymbol{known})$

To estimate the mean of some characteristic or event in a population:

$$estimate\ \pm Z_{\frac{\alpha}{2}} \times\ SE$$

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Where

$\bar{x}$ is the sample mean,

n is the sample size,

σ is the standard deviation of the population,

$Z_{\frac{\alpha}{2}}$ is the point on the standard normal curve area beyond which is $\frac{\alpha}{2}$ %

$1 - \alpha$ is the confidence level.

Example:

One of the properties of a good quality paper is its bursting strength. Suppose a sample of 16 specimens yields mean bursting strength of 25 units, and it is known from the history of such tests that the standard deviation among specimens is 5 units. Assuming normality of test results, what are the (i) 95% and (ii) 98% confidence limits for the mean bursting strength from this sample?

Solution:

$$Here\ n = 16, \bar{x} = 25, \sigma = 5$$

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

(i) $Z_{\frac{\alpha}{2}} = 1.96\ at$ 95% confidence level

$$25 \pm 1.96 \frac{5}{\sqrt{16}} = 25 \pm 2.45$$

$$= 22.55, 27.45$$

$The\ 95\%\ confidence\ limits\ for\ the\ mean\ bursting\ strength\ are$ $(22.55, 27.45)$

(ii) $Z_{\frac{\alpha}{2}} = 2.327\ at$ 98% confidence level

$$25 \pm 2.327 \frac{5}{\sqrt{16}} = 25 \pm 2.91$$

$$= 22.09, 27.91$$

$The\ 98\%\ confidence\ limits\ for\ the\ mean\ bursting\ strength\ are$ $(22.09, 27.91)$

Confidence interval of mean $(\boldsymbol{\sigma}\ \mathbf{unknown})$

To estimate the mean of some characteristic or event in a population:

$$estimate\ \pm t_{\frac{\alpha}{2}, n-1} \times SE$$

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \qquad where\ s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Example:

An article in the journal of testing and evaluation presents the following 20 measurements on residual flame time (in seconds) of treated specimen of children's night wear:

9.85	**9.93**	**9.75**	**9.77**	**9.67**
9.87	**9.67**	**9.94**	**9.85**	**9.75**
9.83	**9.92**	**9.74**	**9.99**	**9.88**
9.95	**9.95**	**9.93**	**9.92**	**9.89**

Assume that residual flame follows normal distribution find 95% confidence interval on the mean residual flame time.

Solution: Here $n = 20, 1 - \alpha = 0.95, \alpha = 0.05$ $(\sigma\ unknown)$

$\boldsymbol{x}$	$\boldsymbol{x - \bar{x}}$	$(\boldsymbol{x - \bar{x}})^2$
9.85	0	0
9.87	0.02	0.0004
9.83	-0.02	0.0004
9.95	0.1	0.01
9.93	0.08	0.0064
9.67	-0.18	0.0324
9.92	0.07	0.0049
9.95	0.1	0.01
9.75	-0.1	0.01
9.94	0.09	0.0081
9.74	-0.11	0.0121
9.93	0.08	0.0064
9.77	-0.08	0.0064
9.85	0	0
9.99	0.14	0.0196
9.92	0.07	0.0049
9.67	-0.18	0.0324
9.75	-0.1	0.01
9.88	0.03	0.0009
9.89	0.04	0.0016
197.05		0.1769
$\bar{x} = 9.85$		

$$\sum x_i = 197.05\,, \qquad \sum (x_i - \bar{x})^2 = 0.1769$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{197.05}{20} = 9.85$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{0.1769}{19} = 0.0093$$

Confidence interval is given by

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$$

$$= 9.85 \pm\ 2.309\ \frac{\sqrt{0.0093}}{\sqrt{20}}$$
$$= 9.85 \pm\ .0498$$
$$= 9.8002, 9.8998$$

Therefore the 95% confidence interval is $(9.8002, 9.8998)$.

Confidence Intervals of Proportions:

To estimate the proportion of some characteristic or event in a population:

Select a random sample of size n, count those with event x, the sample proportion is $x/n\ = p$

The sample proportion $\hat{p}$ estimates the population proportion p.

Confidence Intervals of Proportions for large samples:

Confidence interval formula for the true proportion, p provided $n\hat{p} \geq 5\ \&\ n\hat{q} \geq 5$ (large sample size)

$$estimate\ \pm Z_{\frac{\alpha}{2}} \times\ SE$$

$$\hat{p}\ \pm\ Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\ \ \hat{q}}{n}}\ where\ \ \hat{q} = 1 - \hat{p}$$

Confidence Intervals of Proportions for small samples:

Confidence interval formula for the true proportion, p provided $n\hat{p} < 5\ \&\ n\hat{q} < 5$ (small sample size)

$$estimate\ \pm t_{\frac{\alpha}{2},n-1} \times\ SE$$

$$\hat{p}\ \pm\ t_{\frac{\alpha}{2},n-1}\sqrt{\frac{\hat{p}\ \ \hat{q}}{n}}\ where\ \ \hat{q} = 1 - \hat{p}$$

Example:

An insurance company policy sells foreign travel policy to those going abroad. The company in reputed to settle the claims within a period of 2 months. However, the new CEO of the company came to know about the delay in settling claims. He, therefore, ordered the concerned official to take a sample of 100 claims, and report the proportion of cases which were settled within 2 months. The CEO received the proportion as 0.6. What are the 95% confidence limits for such proportion?

Solution:

$Here\ n = 100, \hat{p} = 0.6$

$$\hat{p}\ \pm\ Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\ \ \hat{q}}{n}}\ \ where\ \ \hat{q} = 1 - \hat{p}$$

$Z_{\frac{\alpha}{2}} = 1.96\ at$ 95% confidence level

$$0.6\ \pm 1.96\ \sqrt{\frac{0.6 \times 0.4}{100}}$$

$$= 0.6\ \pm 0.096\ = 0.504, 0.696$$

Thus, the CEO could be 95% confident that about 50 t0 70% of claims are settled within two months.

Example:

A national television network samples 1400 voters after each has cast a vote in a state gubernatorial election. Of these 1400 voters, 742 claim to have voted for the Democratic candidate and 658 for the Republican candidate. There are only two candidates in the election.

i) Assuming that each sampled voter actually voted as claimed and that the sample is a random sample from the population of all voters is there enough evidence to predict the winner of the election? Base your decision on a 95% confidence interval.

ii) Base your decision on a 99% confidence interval. Explain why it requires greater evidence to make a prediction when we require greater confidence of being correct.

Solution:

$$Here\ n = 1400, \hat{p} = \frac{x}{n} = \frac{742}{1400} = 0.53$$

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\ \hat{q}}{n}}\ \ where\ \ \hat{q} = 1 - \hat{p}$$

$i)\ Z_{\frac{\alpha}{2}} = 1.96\ at$ 95% confidence level

$$0.53 \pm 1.96 \sqrt{\frac{0.53 \times 0.47}{1400}}$$

$$= 0.53 \pm 0.026 = 0.504, 0.556$$

Since our interval is over .50, we can be reasonably confident that the Democratic candidate is the winner.

$i)\ Z_{\frac{\alpha}{2}} = 2.58\ at$ 99% confidence level

$$0.53 \pm 2.58 \sqrt{\frac{0.53 \times 0.47}{1400}}$$

$$= 0.53 \pm 0.034 = 0.496, 0.564$$

Since our interval includes .5 (and goes slightly below), we cannot confidently determine a winner based on this sample.

Confidence interval for the difference between the two means$(\boldsymbol{\mu_1 - \mu_2})$:

$$estimate \pm Z_{\frac{\alpha}{2}} \times SE$$

$$(\overline{x_1} - \overline{x_2}) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Example:

According to an automobile agency, average lives of two premium brands, 'A' and 'B' of car tyres are 45,000 kms and 42,000 kms. Suppose that these mean lives are based on random samples of 50 brand 'A' and 40 brand 'B', and that the standard deviations of these two brands were 3000 kms and 2000 kms, respectively.

i) what is the point estimate of the difference in mean lives of the two tyres?

ii) Construct a 95% confidence interval for difference between the two means.

Solution: Let m_1 and m_2 are the population means of two premier brands of tyres 'A' and 'B' respectively.

$$Here\ \overline{x_1} = 45000\ \overline{x_2} = 42000, n_1 = 50, n_2 = 40, \sigma_1 = 3000, \sigma_2 = 2000,$$

$$Z_{\frac{\alpha}{2}}\ at\ 95\%\ confidence\ level = 1.96$$

95% confidence interval for the difference between two means is given by

$$(\overline{x_1} - \overline{x_2})\ \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= (45000 - 42000)\ \pm\ 1.96\ \sqrt{\frac{3000^2}{50} + \frac{2000^2}{40}}$$

$$= 3000\ \pm\ 1.96\ \ (529.15) = 3000\ \pm\ 1037.13$$

$$= 1962.87\ , 4037.13$$

Therefore 95% confidence interval for the difference between two premier brands of tyres is 1962.87 to 4037.13.

Confidence interval for the difference in population proportions $(\boldsymbol{p_1 - p_2})$*:*

$$estimate\ \pm Z_{\frac{\alpha}{2}} \times\ SE$$

$$(\widehat{p_1} - \widehat{p_2})\ \pm\ Z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p_1}\ \widehat{q_1}}{n_1} + \frac{\widehat{p_2}\ \widehat{q_2}}{n_2}}\ where\ \widehat{q_1} = 1 - \widehat{p_1}, \widehat{q_2} = 1 - \widehat{p_2}$$

Example:

A soap manufacturing company wanted to estimate the difference between the proportions of loyal users of its soap in urban and rural areas. In a sample of 1200 users from urban areas, 300 users were found to be loyal users, and in the sample of 1500 from rural areas, 300 were found to be loyal. What

is the point estimate of the difference of proportion of loyal users of the soap in urban and rural areas, Construct a 95% confidence interval for the proportion of the same.

Solution:

Let p_1 and p_2 be the proportion of loyal users in urban and rural areas respectively.

Here $X_1 = 300, X_2 = 300, n_1 = 1200, n_2 = 1500$, $Z_{\frac{\alpha}{2}}\ at\ 95\%\ confidence\ level = 1.96$

$$\widehat{p_1} = \frac{X_1}{n_1} = \frac{300}{1200} = 0.25$$

$$\widehat{p_2} = \frac{X_2}{n_2} = \frac{300}{1500} = 0.20$$

$$\widehat{q_1} = 1 - \ \widehat{p_1}\ = 1 - 0.25 = 0.75$$

$$\widehat{q_2} = 1 - \ \widehat{p_2} = 1 - 0.20 = 0.80$$

The point estimate of the difference of proportion of loyal users of the soap in urban and rural areas is

$$\widehat{p_1} - \widehat{p_2} = 0.25 - 0.20 = 0.05$$

95% confidence interval for the difference between two proportions is given by

$$(\widehat{p_1} - \widehat{p_2})\ \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p_1}\ \widehat{q_1}}{n_1} + \frac{\widehat{p_2}\ \widehat{q_2}}{n_2}}\ where\ \widehat{q_1} = 1 - \ \widehat{p_1}, \widehat{q_2} = 1 - \ \widehat{p_2}$$

$$= (0.25 - 0.20)\ \pm\ 1.96 \sqrt{\frac{(0.25)(0.75)}{1200} + \frac{(0.20)(0.80)}{1500}}$$

$$= 0.05\ \pm\ 1.96\ (\ 0.0162) = 0.05 \pm\ 0.032$$

$$= 0.018, 0.082$$

Therefore 95% confidence interval of the difference of proportion of loyal users of the soap in urban and rural areas is (0.018, 0.082).

Sample size required for estimating mean:

$$P\{|\bar{x} - \ \mu| \leq margin\} = 1 - \alpha$$

$Where\ \ \bar{x}\ is\ the\ sample\ mean$

$\mu\ is\ the\ population\ \ mean,$

$1 - \alpha\ is\ the\ level\ of\ confidence$

Example:

A company wants to determine the average time to complete a certain job. The past records show that the s.d of the completion times for all the workers in the company has been 10 days, and there is no reason to believe that this would have changed. However, the company feels that because of the procedural changes, the mean would have changed. Determine the sample size so that the company may be 95% confident that the sample average remains within ±2 days of the population mean.

Solution:

Given that the company may be 95% confident that the sample average remains within ± 2 days of the population mean and $\sigma = 10$.

i.e., $|\bar{x} - \mu| < 2$

$$P(|\bar{x} - \mu| < 2) = 0.95$$

$$P\left(\frac{|\bar{x} - \mu|}{{}^{\sigma}/_{\sqrt{n}}} < \frac{2}{{}^{\sigma}/_{\sqrt{n}}}\right) = 0.95$$

$$P\left(|z| < \frac{2}{{}^{\sigma}/_{\sqrt{n}}}\right) = 0.95 \ldots (1) \qquad \left[where\ z = \frac{\bar{x} - \mu}{{}^{\sigma}/_{\sqrt{n}}}\right]$$

We know that $P(|z| < 1.96) = 0.95 \ldots (2)$

From (1) and (2), we get

$$\frac{2}{{}^{\sigma}/_{\sqrt{n}}} = 1.96$$

$$\frac{2\sqrt{n}}{\sigma} = 1.96$$

$\frac{2\sqrt{n}}{10} = 1.96 \Rightarrow \sqrt{n} = 1.96 \times 5 \Rightarrow \sqrt{n} = 9.8$

$n = (9.8)^2 = 96.04$

Approximately the sample size is 97.

Sample size required for estimating proportion:

$$P\{|\hat{p} - p_0| \leq margin\} = 1 - \alpha$$

$Where\ \ \hat{p}\ is\ the\ estimated\ proportion$

$p_0\ is\ the\ true\ proportion,$

$1 - \alpha\ is\ the\ level\ of\ confidence$

Example:

20% of the population of a town is supposed to be rice eaters. At 95% level of confidence, what should be the sample size, so that the sampling error is not more than 5% above or below the true proportion of rice eaters?

Solution:

Let p_0 be the proportion of rice eaters in a town.

$$Given\ p_0 = 20\% = 0.2, \alpha = 5\%\ , |\hat{p} - p_0| \leq 0.05$$

$$q_0 = 1 - p_0 = 1 - 0.2 = 0.8$$

$$P(|x - m| \leq 0.05) = 0.95$$

$$P\left(\frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0 q_0}{n}}} \leq \frac{0.05}{\sqrt{\frac{p_0 q_0}{n}}}\right) = 0.95$$

$$P\left(|z| \leq \frac{0.05}{\sqrt{\frac{p_0 q_0}{n}}}\right) = 0.95\ where\ z = \frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0 q_0}{n}}}$$

$$\frac{0.05}{\sqrt{\frac{p_0 q_0}{n}}} = 1.96\ \left[\ Since\ P(|z| \leq 1.96) = 0.95\ \right]$$

$$\frac{0.05}{\sqrt{\frac{0.2 \times 0.8}{n}}} = 1.96 \Rightarrow \frac{0.05\sqrt{n}}{\sqrt{0.16}} = 1.96$$

$$\sqrt{n} = \frac{1.96 \times \sqrt{0.16}}{0.05} = 15.68$$

$$n = 15.68^2 = 245.86$$

Therefore the sample size is approximately 246.