ADM College for Women (Autonomous)

(Accredited with 'A' Grade by NAAC 4th Cycle)
(Affiliated to Bharathidasan University, Thiruchirappalli)
Nagapattinam – 611 001

Department : Economics

Course Name : Statistical Methods for Economic Analysis

Class : I MA

Submitted by : Dr. P. Sujatha

STATISTICAL METHODS FOR ECONOMIC ANALYSIS

UNIT I MEASURES OF AVERAGES AND DISPERSIONS

Measures of Averages, Definition, Characteristics of a good Measure of Average - Mean, Median and Mode - Definition, Merits and Demerits (Simple Problems). Measures of Dispersions - Definition, Characteristics of a good measure of dispersion - Standard Deviation - Definition, Merits and Demerits, Coefficient of Variation, (Simple Problems)

UNIT II CORRELATION ANALYSIS

Correlation Analysis- Definition, Types, Methods of Finding Correlation Co-efficient - Scatter Diagram, Karl Pearson, Spearman's Rank Correlation Co-efficient, Concurrent Deviations Method -Properties of Correlation Co-efficient.(Only statement Without proof) (Simple Problems)

UNIT III REGRESSION ANALYSIS

Regression Analysis –Definition, Types, Regression Equation X on Y and Regression Equation Y on X - Properties of Regression Co- Efficient- (Without proof) - Difference Between Correlation and Regression Analysis. (Simple Problems)

UNIT IV SAMPLING METHODS

Sampling- Definition, Uses of sampling. Random Sampling- Simple Random Sampling Stratified Random Sampling, Systematic Random Sampling - Definition, Merits and Demerits, Non-random sampling - Purposive, Quota and Judgement sampling. (Only Theory).

UNIT V TESTING OF HYPOTHESIS

Sampling distribution of Means, Standard Error- Uses of Standard Error - Testing of Hypothesis - Test Procedure - Type I error, Types II error - One Tailed & Two Tailed Tests, - t - test- Testing Significance of Single Mean and Difference between Two Means, Chi square test-Testing the Independence of Two Attributes, (Simple Problems).

Text Books:

1. Gupta S.P - Statistical methods, Sultan Chand and Son's New Delhi, 2014,
2. Gupta, S.C- Fundamentals of Applied Statistics, Sultan Chand and son's New Delhi, 2005.

UNIT I MEASURES OF AVERAGES AND DISPERSIONS

Definition:

Data show a tendency to concentrate at certain values, usually somewhere in the centre of the distribution. Measures of this tendency are called measures of central tendency.

Average:

Average is a value which is typical or representative of a set of data.

Objectives of averages:

1. Average provides a quick understanding of complex data.
2. Averages enable comparison.
3. Averages facilitate sampling techniques.
4. Averages give the way for further statistical analysis.
5. Averages establish the relationship between variables.

Different Measures of Central Tendency:

1. Arithmetic Mean (or) Mean
2. Median
3. Mode
4. Geometric Mean
5. Harmonic Mean

1. Arithmetic Mean (A.M):

Arithmetic Mean is the total of values of the items divided by their number. It is denoted by $\bar{X}$.

Methods of finding Mean:

Direct Method:

1. $\bar{X} = \frac{\Sigma X}{N}$; N= number of observations (Individual Observations)
   X = value
2. $\bar{X} = \frac{\Sigma f X}{N}$; X=value, N = $\sum$ f, (Discrete frequency distribution)
   f = frequency
3. $\bar{X} = \frac{\Sigma f m}{N}$; N = $\sum$ f, (Continuous frequency distribution)
   f = frequency, m = mid values

Short-Cut Method:

1. Individual Observations

$\bar{X}$ = A ± $\frac{\Sigma d}{N}$, A = assumed mean, d = X-A,
N = number of values

2. Discrete frequency distribution

$\bar{X}$= A ± $\frac{\Sigma fd}{N}$, A = assumed mean, f = frequency
d = X-A , N = ∑ f = Total Frequency

3. Continuous frequency distribution

$\bar{X}$= A ± $\frac{\Sigma fd}{N}$ X c

Here A = Assumed Mean,
N = Σf = Total Frequency,
d = $\frac{m - A}{c}$
m = middle value of class interval,
c = class interval

Example 1:

Calculate the mean from the following data:

Marks	40	50	55	78	58	60	73	35	43	48

Solution:

∑X = Sum of Marks = 540

Mean = $\bar{X}$= $\frac{\Sigma X}{N}$ = $\frac{540}{10}$
= 54 Marks

Example 2:

Calculate Mean from the following data:

Value	1	2	3	4	5	6	7	8	9	10
Frequency	21	30	28	40	26	34	40	9	15	57

X	f	fX
1	21	21
2	30	60
3	28	84
4	40	160
5	26	130
6	34	204
7	40	280
8	9	72
9	15	135
10	57	570
	$\sum$ f = 300	$\sum$ f X = 1716

Mean = $\bar{X} = \frac{\sum f\ X}{N} = \frac{1716}{300}$

= 5.72

Example 3:

Find out the mean profits:

Profits per shop Rs.	Number of shops
100 - 200	10
200 - 300	18
300 - 400	20
400 - 500	26
500 - 600	30
600 - 700	28
700 - 800	18

Solution:

(Please refer below table for computations.)

Mean = $\bar{X} = \frac{\sum fm}{N}$

$= \frac{72900}{150}$

= 486

∴Average Profit is **Rs.486.**

Profits Rs.	Mid point (m)	No. of shops (f)	fm
100 - 200	150	10	1500
200 - 300	250	18	4500
300 - 400	350	20	7000
400 - 500	450	26	11700
500 - 600	550	30	16500
600 - 700	650	28	18200
700 - 800	750	18	13500
		∑f = 150	∑f m = 72900

Properties of Good Average:

1. It should be rigidly defined.
2. It should be based on all the items.
3. It should not be unduly affected by extreme items.
4. It should lend itself for algebraic manipulations.
5. It should be simple to understand and easy to calculate.
6. It should have sampling stability.

Merits of Arithmetic Mean:

1. It is rigidly defined.
2. It is based on all the items.
3. It lends itself for algebraic manipulations.
4. It is simple to understand and easy to calculate.
5. It has sampling stability.
6. It is the most useful measure of central tendency.
7. Arithmetic mean is the best measure of central tendency.

Demerits of Arithmetic Mean:

1. It is affected by extreme items.
2. It cannot be calculated for open-end data.
3. It cannot be found graphically.
4. It cannot be found for qualities.
5. It cannot be calculated even when one of the values is missing.

2. Median:

Median is the middle most value when all the observations are in ascending or descending order. It is denoted by M.

Median divides the whole distribution into two equal parts, one half containing values greater than it and the other half containing values less than it. Therefore, the values have to be arranged in ascending or descending order, before finding the median.

As distinct from the mean which is calculated from the value of every item in the series, the median is found based on the position of a value. Therefore, median is called positional average.

Methods of finding Median:

Individual Observations:

M = Size of $\left(\frac{n+1}{2}\right)^{th}$item; n = Number of observations

Discrete Frequency Distribution:

M = Size of $\left(\frac{N+1}{2}\right)^{th}$ item; N = ∑f

Continuous Frequency Distribution:

M = $L + \frac{\frac{N}{2} - cf}{f} Xh$

Where L = lower limit of the median class

c.f = cumulative frequency preceding the median class

f = frequency of the median class

h = size of the median class

Example 1:

The following are the marks scored by 7 students; find out the median marks:

Marks	*45*	*32*	*18*	*57*	*65*	*28*	*46*

Solution:

Computation of Median Marks: 1st arrange the data in Ascending Order

Marks	*18*	*28*	*32*	*45*	*57*	*58*	*65*

Median = Value in $\left(\frac{n+1}{2}\right)$ th item

$= \frac{7+1}{2}$th item

$= \frac{8}{2}$th item

$= 4^{th}$ item

= 45

Therefore, the median mark is 45.

Example 2: Find out the median from the following:

Values	*62*	*71*	*57*	*58*	*61*	*42*	*38*	*65*	*72*	*66*

Solution:

Values	*38*	*42*	*57*	*58*	*61*	*62*	*65*	*66*	*71*	*72*

Median = Value in $\left(\frac{n+1}{2}\right)$ th value

$= \frac{10+1}{2}$th value

= 5.5 th value

Median = $\left(\frac{5th\ value + 6th\ value}{2}\right)$

$= \left(\frac{61+62}{2}\right)$

$= \left(\frac{123}{2}\right)$

= 61.5

Example 3:

Locate the median from the following:

Size of Shoes	*5*	*5.5*	*6*	*6.5*	*7*	*7.5*	*8*
Frequency	*10*	*16*	*28*	*15*	*30*	*40*	*34*

Solution:

Size of Shoes	**f**	**cf**
5	10	10
5.5	16	10+16=26
6	28	26+28=54
6.5	15	54+15=69
7	30	69+30=99
7.5	40	99+40=139
8	34	139+34=173

Median=Size of shoe in $\left(\frac{n+1}{2}\right)$th value

= Size of shoe in $\left(\frac{173+1}{2}\right)$th value

=Size of shoein87th item

=7

Example 4:

Calculate the median from the following table:

Marks	Frequency
10 - 25	6
25 - 40	20
40 - 55	44
55 - 70	26
70 - 85	3
85 - 100	1

Solution:

Marks (x)	Frequency (f)	Cumulative frequency (cf)
10 - 25	6	6
25 - 40	20	26
40 - 55	44	70
55 - 70	26	96
70 - 85	3	99
85 - 100	1	100

Solution:

Median= Mark in $\left(\frac{n}{2}\right)$th value

= Mark in $\left(\frac{100}{2}\right) th$value

= 50^{th} value

Median item lies in 40 – 55marks group.

Median = L + $\frac{\frac{N}{2}-cf}{f}$Xh

Here L = 40, N/2 = 50, cf = 26, f = 44 and h = 15

= 40 + $\frac{50-26}{44}$X15

= 40 + $\frac{24X15}{44}$

= 40 + $\frac{360}{44}$

= 40 + 8.18

= 48.18 marks

3. **Mode:**

Mode is the value which has the greatest frequency density. It is denoted by **Z.**

Mode is the most common item of a series. It is the value which occurs greatest number of times in a series. It is derived from the French word 'La mode' meaning the fashion. Mode is the most fashionable or typical value of the distribution, because it is repeated the highest number of times in the series. Mode is defined as the value of the variable which occurs most frequently in a distribution. The mode in a distribution is that item around which there is a maximum concentration.

Methods of Finding Mode:

Individual Observations:

Z = Value that occurs more times

Discrete Frequency Distribution:

Z = Value with greatest frequency density $2f_1 – f_0 – f_2$

Continuous Frequency Distribution:

$$Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \text{ X } h$$

Where L = lower limit of modal class
f_1 = frequency of the modal class
f_0 = frequency preceding the modal class
f_2 = frequency succeeding the modal class
h= size of the modal class.

Raw Data (or) Individual Series:

Example 1:

Calculate mode from the following data.

Value	*25*	*30*	*40*	*32*	*40*	*50*	*25*	*40*

Solution:

Mode (z) = The value which is repeated maximum number of times

Mode (z)= 40

Discrete Frequency Distribution:

Example 2:

Calculate mode from the following data.

Value	*50*	*52*	*60*	*68*	*70*	*72*	*80*
F	*3*	*7*	*15*	*25*	*20*	*8*	*2*

Solution:

Mode (z) = The value which has highest frequency.

Mode (z)= 68

Continuous Frequency Distribution:

Example 3:

Calculate mode from the following data:

Class Interval	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
Frequency (f)	8	12	20	32	17	3

Solution:

Mode (Z) = $L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

Where L = lower limit of the modal class.

f_1 = frequency of the modal class. f_0= frequency of the preceding class to the modal class.

f_2 = frequency of the succeeding class to the modal class. h = Length of the modal class.

Modal class = The class which has the highest frequency.

= 30 - 40

Therefore, L= 30, $f_1 = 32, f_0 = 20, f_2 = 17, h = 10$

Mode (Z) = $30 + \frac{32-20}{2(32)-20-17} \times 10$

$= 30 + \frac{12}{64-37} \times 10$

$= 30 + \frac{12 \times 10}{27}$

$= 30 + \frac{120}{27}$

$= 30 + 4.44 = 34.44$

Merits of Median and Mode:

1. It is simple to understand.
2. It is easy to calculate.
3. It is suitable for open-end classes.

4. It can be determined graphically.
5. It is not affected by extreme values.
6. It is used to deal with qualitative data.

Demerits of Median and Mode:

1. It is not rigidly defined.
2. It is not based on all the items.
3. It cannot be manipulated algebraically.
4. It does not have sampling stability.
5. It is difficult to find when there are large number of items. (Only for median)
6. It is used lesser than arithmetic mean.

Empirical Formula:

Mode = 3 Median - 2 Mean

Properties of Arithmetic Mean:

i) $\sum(X - X) = 0$

ii) $\sum(X - X)^2$ is least.

Measures of Dispersion

Definition:

Dispersion is the measure of the variations of the items. The degree to which numerical data tend to spread about an average value is called the variation or dispersion of data.

Uses of Dispersion:

1. The reliability of a measure of central tendency is known.
2. Measures of dispersion provide a basis for the control of variability.
3. They help to compare two or more sets of data with regard to their variability.
4. They enhance the utility and scope of desirable properties of a measure of dispersion.

Various Measures of Dispersion:

	Absolute measure	Relative measure
1.	Range	Coefficient of range
2.	Quartile Deviation	Coefficient of Quartile Deviation
3.	Mean Deviation	Coefficient of Mean Deviation
4.	Standard Deviation	Coefficient of Variation
5.	Variance	

Properties of Good Measure of Dispersion:

1. It should be rigidly defined.
2. It should be based on all the items.
3. It should not be unduly affected by extreme items.
4. It should lend itself for algebraic manipulation.
5. It should be simple to understand and easy to collect.
6. It should have sampling stability.

Standard Deviation(σ):

Standard Deviation is the root mean square deviation of values from their arithmetic mean.

1. $\sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \sqrt{\frac{\Sigma X^2}{N} - (\bar{X})^2}$ (for individual observation)
2. $\sqrt{\frac{\Sigma f(X-\bar{X})^2}{N}} = \sqrt{\frac{\Sigma f X^2}{N} - (\bar{X})^2}$ (for frequency observation)

Coefficient of Variation:

Very popular and extensively used relative measure of dispersion.

Coefficient Variation = $\frac{Standard\ \ Deviation}{Mean} \times 100$

Uses of Standard Deviation:

1. Standard Deviation is the best absolute measure of dispersion.
2. It is a part of many statistical concepts such as skewness, kurtosis, etc..
3. It is used in statistical quality control.

Merits of Standard Deviation:

1. Standard Deviation is rigidly defined.
2. It is calculated on the basis of all the items.
3. It could be manipulated further. The combined standard deviation can be calculated.
4. Mistakes in its calculation can be corrected. The entire calculation need not be redone.
5. It is the most important absolute measure of dispersion. It is used in all the areas of statistics.
6. Scientific calculators show the standard deviation of any series.

Demerits of Standard Deviation:

1. Compared to other measures of dispersion, standard deviation is difficult to calculate.
2. It is not simple to understand.
3. It gives more weightage to the items away from the mean than those near the mean as the deviations are squared.

Problem:

Calculate the standard deviation and coefficient of variation from the following data:

X : 14 22 9 15 20 17 12 11

Solution:

Standard Deviation, $\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}}$

Coefficient of Variance = $\frac{Standard\ \ Deviation}{Mean} \times 100$

Calculation of S.D from Actual Mean

Values (X)	$X - \bar{X} = X - 15$	$(X - \bar{X})^2$
14	14 - 15 = -1	-1 x -1 =1
22	22 - 15 = 7	7 x 7 = 49
9	9 - 15 = -6	6 x6 =36
15	15 -15 =0	0x0 =0
20	20 - 15 =5	5x5=25
17	17 - 15 = 2	2x2 =4
12	12 - 15 =-3	-3x-3 =9
11	11 - 15 =-4	-4 x-4=16
$\sum X = 120$		$\sum (X - \bar{X})^2 = 140$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{120}{8} = 15$$

$$\sigma = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}} = \sqrt{\frac{140}{8}} = \sqrt{17.5} = 4.18$$

Coefficient of Variance = $\frac{Standard\ \ Deviation}{Mean} \times 100 = \frac{4.18}{15} \times 100 = \frac{418}{15} = 27.87$

Result: Standard deviation =4.18 Coefficient of variation = 27.87

Alternate Solution for calculating standard deviation:

Values = X	**X^2**
14	14x14=196
22	22 x22 =282
9	9x9 =81
15	15 x15 =225
20	20 x 20=400
17	17 x 17 = 209
12	12x12 =144
11	11 x 11=121
$\sum X = 120$	$\sum X^2 = 1940$

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} = \sqrt{\frac{1940}{8} - \left(\frac{120}{8}\right)^2} = \sqrt{242.5 - 225} = \sqrt{17.5} = 4.18$$

Problem:

Calculate standard deviation and coefficient of variation from the following:

Marks	*No. of Students*
10	*8*
20	*12*
30	*20*
40	*10*
50	*7*
60	*3*

X	f	fX	$(fX)X=fX^2$
10	8	10x8=80	80x10= 800
20	12	20x12=240	240x20=4800
30	20	30x20=600	600x30=18000
40	10	40x 10= 400	400x40=16000
50	7	50x 7 =350	350x50=17500
60	3	60x 3= 180	180x60=10800
	N=60	$\sum fx= 1850$	$\sum fx^2 =67900$

Solution:

Standard Deviation, σ = $\sqrt{\frac{\Sigma fX^2}{N} - (\bar{X})^2}$

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{1850}{60} = 30.83$$

$$\frac{\Sigma fX^2}{N} = \frac{67900}{60} = 1131.67$$

$$\bar{X} \times \bar{X} = 30.83 \times 30.83 = 950.49$$

Standard Deviation, σ= $\sqrt{\frac{\Sigma fX^2}{N} - (\bar{X})^2} = \sqrt{1131.67 - 950.49} = \sqrt{181.18} = 13.46$

Coefficient of Variance = $\frac{Standard\ \ Deviation}{Mean} \times 100 = \frac{13.46}{30.83} \times 100 = \frac{1346}{30.83} = 43.66$

Result:

Standard deviation =13.46 and Coefficient of Variance = 43.66.

Problem:

Prices of a particular commodity in five years in two cities are given below:

Price in city A	**Price in city B**
20	10
22	20
19	18
23	12
16	15

From the above data, find the city which had more stable prices.

Solution:

City A			City B		
Prices(X)	Deviations from (X) $(X-\overline{X})= X-20$	Square of the deviations $=(X-\bar{X})^2$	Prices(Y)	Deviations from Y $(Y-\overline{Y})=Y-15$	Square of the deviations $=(Y-\bar{Y})^2$
20	20 -20=0	0x0=0	10	10- 15=-5	-5x-5=25
22	22 -20=+2	2x2=4	20	20 -15=+5	5x5=25
19	19 -20=-1	-1x-1=1	18	18 -15 =+3	3x3=9
23	23 -20=+3	3x3=9	12	12 -15 =-3	-3x-3=9
16	16 -20=-4	-4x-4=16	15	15-15=0	0x0=0
$\Sigma X = 100$	$\Sigma(X - \overline{X})= 0$	$\Sigma(X-\overline{X})^2= 30$	$\Sigma Y= 75$	$\Sigma (Y-\overline{Y})= 0$	$\Sigma (Y-\overline{Y})^2= 68$

City A:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{100}{5} = 20$$

$$\sigma_x = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}} = \sqrt{\frac{30}{5}} = \sqrt{6} = 2.45$$

$$C.V. = \frac{\sigma_x}{\bar{X}} \times 100 = \frac{2.45}{20} \times 100 = 12.25$$

City B:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{75}{5} = 15$$

$$\sigma_x = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}} = \sqrt{\frac{68}{5}} = \sqrt{13.6} = 3.69$$

$$C.V. = \frac{\sigma_x}{\bar{X}} \times 100 = \frac{3.69}{15} \times 100 = 24.6$$

Result:

C.V. for A= 12.25 and C.V. for B = 24.6.

City A had more stable prices than in city B, because the coefficient of variations is lower in city A.

Variance(σ^2):

Variance is the mean square deviation of the values from their arithmetic mean.

Problem:

Calculate coefficient of variation for the following data:

25 15 23 40 27 25 23 25 20

Solution:

Coefficient of Variance = $\frac{Standard\ \ Deviation}{Mean} \times 100$

Mean, $\bar{X} = \frac{\Sigma X}{N}$Standard Deviation,$\sigma = \sqrt{\frac{\Sigma X^2}{N} - (\bar{X})^2}$

Z = The most repeated value

X	*25*	*15*	*23*	*40*	*27*	*25*	*23*	*25*	*20*	*$\sum X$=223*
X^2	*625*	*225*	*529*	*1600*	*729*	*625*	*529*	*625*	*400*	*$\sum X^2$=5887*

$$\bar{X} = \frac{\Sigma X}{N} = \frac{223}{9} = 24.78$$

Standard Deviation,

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - (\bar{X})^2} = \sqrt{\frac{5887}{9} - (24.78)^2} = \sqrt{654.11 - 614.05} = \sqrt{40.06} = 6.33$$

Coefficient of Variance = $\frac{Standard\ \ Deviation}{Mean} \times 100$

$= \frac{6.33\ x100}{24,78} = \frac{633}{24.78} = 25.54$

Result:

Hence Coefficient of Variance = 25,54,

Problem:

Find the Coefficient of Variation from the data given below:

Class	*10 - 15*	*15 - 20*	*20 - 25*	*25 - 30*	*30 - 35*
Frequency	*2*	*17*	*26*	*20*	*15*

Solution:

$\bar{X} = \frac{\Sigma fm}{N}$, m = mid-point of the class $\sigma = \sqrt{\frac{\Sigma fm^2}{N} - (\bar{X})^2}$

$$\bar{X} = \frac{\Sigma fm}{N} = \frac{1945}{80} = 24.31$$

Class	*f*	*m*	*fm*	*(fm)m=fm²*
10 – 15	*2*	*12.5*	*2x12.5=25*	*25x12.5=312.5*
15 – 20	*17(f_0)*	*17.5*	*17x17.5=297.5*	*297.5x17.5=5206.25*
20 – 25	*26(f_1)*	*22.5*	*26x22.5=585*	*585x22.5=13162.5*
25 – 30	*20(f_2)*	*27.5*	*20x27.5=550*	*550x27.5=15125*
30 - 35	*15*	*32.5*	*15x32.5=487.5*	*487.5x32.5=15843.75*
	N=80		*Σfm =1945*	*∑fm²=49650*

$$\sigma = \sqrt{\frac{\Sigma f m^2}{N} - (\bar{X})^2} = \sqrt{\frac{49650}{80} - (24.31)^2} = \sqrt{620.63 - 590.98} = \sqrt{29.65} = 5.45$$

Coefficient of Variance = $\frac{Standard\ \ Deviation}{Mean} \times 100$

$= \frac{5.45\ x100}{24{,}31} = \frac{545}{24.31} = 24.418 = 24.42$

Result:

Hence Coefficient of Variance = 24.42

UNIT II CORRELATION ANALYSIS

Introduction

Sometimes it may happen that values of the variables so collected are interrelated. We may be interested to find if there is any relationship between the two variables under study. For example: the price of the commodity and its sale, with increase in the prize of the product, the quantity sold is bound to decrease, or with decrease in the price of the product the quantity sold is bound to increase. Therefore, we conclude that some relationship between price and sale.

Correlation is the statistical analysis which measures the closeness of the relationship between the variables. The word relationship is of important and indicates that there is some connection between variables under observation.

Uses of Correlation:

i. Correlation is very useful to economists to study the relationship between variables, like price and quantity demanded.For businessmen, it helps to estimate costs, sales, price and other related variables.
ii. Correlation analysis helps in measuring the degree of relationship between the variables.
iii. The relation between variables can be verified and tested for significance, with the help of the correlation analysis. The effect of correlation is to reduce the range of uncertainty of our prediction.
iv. The coefficient of correlation is a relative measure and we can compare the relationship between variables.

Definition:

Correlation refers to the relationship between two or more variables.

For example: price and demand, yield of crops and rainfall.

Types of Correlation:

1. Positive correlation and negative correlation.
2. Simple, multiple, partial correlation.
3. Linear and nonlinear correlation.

Positive Correlation

When the values of two variables change in the same direction, there

is positive correlation between the two variables.

For example: height and weight, rainfall and yield of crops, price and supply

Negative Correlation

When the values of two variables change in the opposite direction, there is negative correlation between the two variables.

For example: price and demand.

Simple Correlation

When we study only two variables, the relationship is described as simple correlation.

For example: demand and price, price and supply, etc.

Multiple Correlation

When we study more than two variables simultaneously, the relationship is described as multiple correlation.

For example: relationship between price, demand and supply of a commodity

Partial Correlation

When more than two variables are considered, the correlation between two of them when all other variables are held constant is called partial correlation.For example: When we study price and demand , eliminating the supply side.

Linear Correlation

The ratio of change between two variables is uniform, then there will be linear correlation between them. Consider the following

X	5	10	15	20
Y	4	8	12	16

The ratio of change between two variables is same. If we plot these points on the graph, we get a straight line.

Non-Linear Correlation

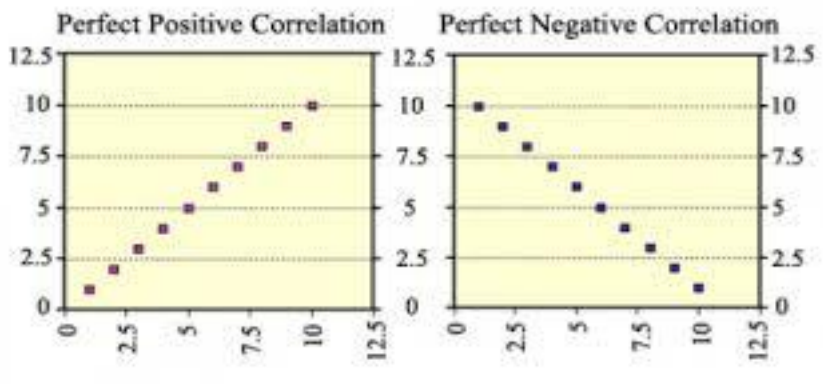
In a curvilinear or nonlinear correlation, the amount of change in one variable does not bear a constant ratio of the amount of change in the other variables.

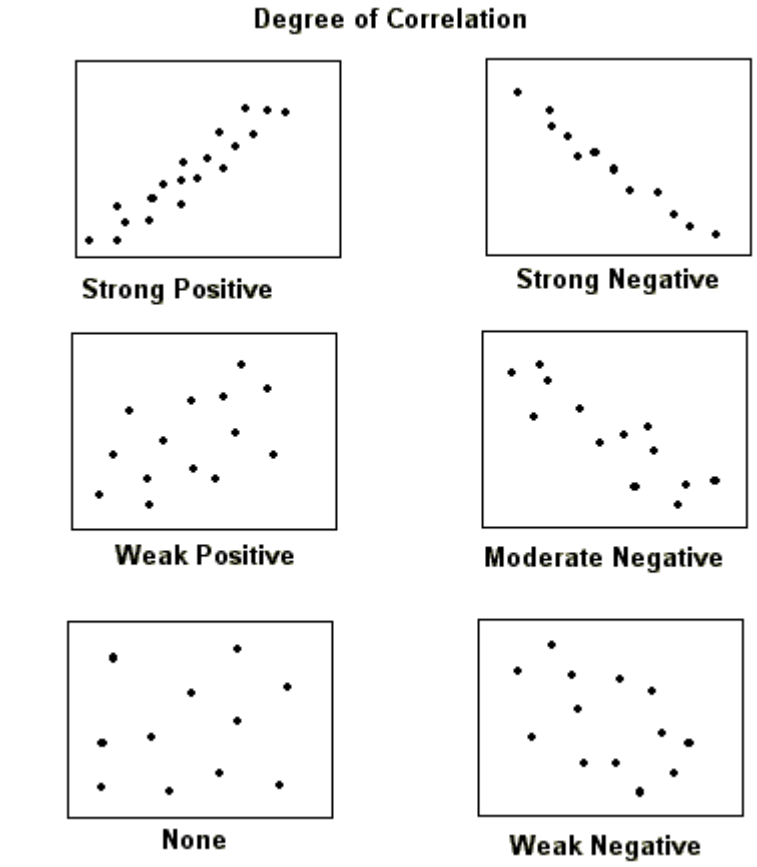
Methods for Studying Correlation

1. Scatter Diagram
2. Karl Pearson’s Correlation coefficient method
3. Spearman’s rank correlation method
4. Concurrent deviation method

Scatter Diagram

Let (X_i , Y_i) i = 1,2,3...N be the pairs of values of two variables X and Y. A point is plotted on a graph sheet corresponding to each pair of the values. The resulting diagram with N points is called Scatter Diagram. Possible types of scatter diagram under simple linear correlation are:





Merits of Scatter Diagram

1. It is easy to draw.
2. It is non-mathematical method.
3. It is simple to understand.
4. This does not involve computations.

Demerits of Scatter Diagram

1. It gives only a rough idea.
2. Comparison is not possible.

Karl Pearson's Correlation Coefficient

Karl Pearson's correlation coefficient is a mathematical the magnitude of linear relationship between two variables. It is the most widely used method in practice and is known as Pearsonian coefficient of correlation. It is denoted by 'r' and is defined as:

$$r = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \quad \text{OR} \quad r = \frac{\Sigma xy}{N\sigma_X \sigma_Y}$$

Where

$$x = X - \bar{X}$$
$$y = Y - \bar{Y}$$
$$COV(X,Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N}$$

$$\sigma_X = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} \qquad \sigma_Y = \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{N}}$$

$$r = \frac{(N\Sigma XY) - ((\Sigma X)(\Sigma Y))}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y - (\Sigma Y)^2]}}$$

Range of coefficient of correlation

The value of the coefficient of correlation lies between -1 and +1.

Interpretation of coefficient of correlation

When r= +1, then there is a perfect positive correlation between the variables. When r = -1, then there is a perfect negative correlation between the variables. When r = 0, then there is no relationship between the variables.

Properties of Correlation coefficient

i. The correlation coefficient lies between -1 to +1 i.e., $-1 \leq r \leq +1$.

ii. The correlation coefficient is independent of change of origin and scale.

iii. The coefficient of correlation is the geometric mean of two regression coefficients.

Merits and Demerits of Karl Pearson's Correlation Coefficient Method:

Merits:

i. Karl Pearson's method is the most popular mathematical method.

ii. The coefficient of correlation summarized in one figure the degree of relation and its direction.

Demerits:

i. Assumption of linear relationship between variables is not affected, whether it is correct or not.

ii. The calculation of coefficient of correlation is time consuming.

Probable Error:

The probable error of the coefficient of correlation is defined as follows:

$$P.E_r = 0.6745\frac{1-r^2}{\sqrt{N}}$$

If the value of **r** is more than six times the probable error, the value of **r** is significant.

Problem 1:

Calculate Karl Pearson's coefficient of correlation from the following data.

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	3

Solution:

Karl Pearson's coefficient of correlation (r)=$\frac{N\Sigma XY-\Sigma X\Sigma Y}{\sqrt{[N\Sigma X^2-(\Sigma X)^2]\times[N\Sigma Y^2-(\Sigma Y)^2]}}$

Computation of Coefficient of Correlation

X	**Y**	**X^2**	**Y^2**	**XY**
12	14	144	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	3	49	9	21
$\Sigma X = 70$	$\Sigma Y = 63$	$\Sigma X^2 = 728$	$\Sigma Y^2 = 651$	$\Sigma XY = 676$

$\Sigma XY = 676$ $\Sigma X = 70$ $\Sigma Y = 63$ $\Sigma X^2 = 728$ $\Sigma Y^2 = 651$ $N = 7$

$$r = \frac{(N \times \Sigma XY) - [(\Sigma X) \times (\Sigma Y)]}{\sqrt{[(N \times \Sigma X^2) - (\Sigma X)^2] \times [(N \times \Sigma Y^2) - (\Sigma Y)^2]}}$$

$$r = \frac{(7 \times 676) - (70 \times 63)}{\sqrt{[(7 \times 728) - (70)^2] \times [(7 \times 651) - (63)^2]}}$$

$$= \frac{4732 - 4410}{\sqrt{(5096 - 4900) \times (4557 - 3969)}} = \frac{322}{\sqrt{196 \times 588}} = \frac{322}{339.48}$$

$$= +0.95$$

Result :

Karl Pearson's coefficient of correlation = 0.95.

Hence the given data are positively correlated.

Problem 2:

Find Karl Pearson's coefficient of correlation between the heights and weights given below:

Height in inches	57	59	62	63	64	65	55	58	57
Weight in lbs.	113	117	126	126	130	129	111	116	112

Solution:

Karl Pearson's coefficient of correlation

$$(r) = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N\sigma_x\sigma_y}$$

X Series:

$\Sigma X=540$ $N=9$ $\Sigma(X - \bar{X})^2 = 102$

Arithmetic Mean,

$$\bar{X} = \frac{\Sigma X}{N} = \frac{540}{9} = 60$$

Standard Deviation,

$$\sigma_x = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \sqrt{\frac{102}{9}} = \sqrt{11.33} = 3.36$$

Y Series:

ΣY=1080 N=9 $\Sigma(Y-\bar{Y})^2 = 472$ $\Sigma(X-\bar{X})(Y-\bar{Y}) = 216$

Arithmetic Mean,

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{1080}{9} = 120$$

Standard Deviation,

$$\sigma_y = \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{N}} = \sqrt{\frac{472}{9}} = \sqrt{52.44} = 7.241$$

X Series			Y Series			Product of dev. of X and Y series $(X-\bar{X})(Y-\bar{Y})$
Height in inches X	Dev. from Mean= 60 $(X-\bar{X})$	Square of dev. $(X-\bar{X})^2$	Weight in lbs, Y	Dev. from Mean=120 $(Y-\bar{Y})$	Square of dev. $(Y-\bar{Y})^2$	
57	57 -60=-3	-3x-3=9	113	113-120=-7	-7x-7=49	-3x-7=21
59	59-60=-1	-1x-1=1	117	117-120=-3	-3x-3=9	-1x-3=3
62	62-60=+2	2x2=4	126	126-120=+6	6x6=36	2x6=12
63	63-60=+3	3x3=9	126	126-120=+6	6x6=36	3x6=18
63	64-60=+4	4x4=16	130	130-120=+10	10x10=100	4x10=40
65	65-60=+5	5x5=25	129	129-120=+9	9x9=81	5x9=45
55	55-60=-5	-5x-5=25	111	111-120=-9	-9x9=81	-5x-9=45
58	58-60=-2	-2x-2=4	116	116-120=-4	-4x-4=16	-2x-4=8
57	57-60=-3	-3x-3=9	112	112-120=-8	-8x-8=64	-3x-8=24
ΣX=540	$\sum(X-\bar{X})=0$	$\sum(X-\bar{X})^2=102$	ΣY=1080	$\sum(Y-\bar{Y})=0$	$\sum(Y-\bar{Y})^2=472$	$\sum(X-\bar{X})(Y-\bar{Y})=216$

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N \times \sigma_x \times \sigma_y} = \frac{216}{9 \times 3.36 \times 7.241} = +0.98$$

Alternate Solution:

$$(r) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2 \times (Y-\bar{Y})^2}} = \frac{216}{\sqrt{102 \times 472}} = \frac{216}{\sqrt{48144}} = \frac{216}{219.42} = 0.9844$$

r = 0.98

Result

Karl Pearson's coefficient of correlation= 0.98.

Hence the given data are positively correlated.

Problem 3:

Find Karl Pearson's coefficient of correlation between the heights and weights given below:

X	100	110	120	105	125	130
Y	90	85	80	75	72	70

Solution:

X	$(X-\bar{X})$	$(X-\bar{X})^2$	Y	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
100	-15	225	90	+11.33	128.37	-169.95
110	-5	25	85	+6.33	40.07	-31.65
120	+5	25	80	+1.33	1.77	6.65
105	-10	100	75	-3.67	13.47	36.7
125	+10	100	72	-6.67	44.49	-66.7
130	+15	225	70	-8.67	75.17	-130.05
$\sum X=690$	$\sum(X-\bar{X})=0$	$\sum(X-\bar{X})^2=700$	$\sum Y=472$	$\sum(Y-\bar{Y})=0$	$\sum(Y-\bar{Y})^2=303.34$	$\sum(X-\bar{X})(Y-\bar{Y})=-355$

X Series:

Mean of X,

$$\bar{X} = \frac{\Sigma X}{N} = \frac{690}{6} = 115 \qquad \Sigma(X-\bar{X})^2 = 700$$

Y Series:

Mean of Y,

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{472}{6} = 78.67 \qquad \Sigma(Y-\bar{Y})^2 = 300.34$$

$$\Sigma(X-\bar{X})(Y-\bar{Y}) = -355$$

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2(Y-\bar{Y})^2}} = \frac{-355}{\sqrt{700 \times 300.34}} = \frac{-355}{212338} = \frac{-355}{460.80} = -0.7703$$

r = -0.77

Result:

Coefficient of correlation(r) is -0.77. Hence there is negative correlation.

Problem 4:

Calculate Karl Pearson's coefficient of correlation from the following data.

X	10	12	14	16	17
Y	5	6	8	9	10

Solution:

Karl Pearson's coefficient of correlation

$$(r) = \frac{(N \times \Sigma XY) - [(\Sigma X) \times (\Sigma Y)]}{\sqrt{[(N \times \Sigma X^2) - (\Sigma X)^2] \times [(N \times \Sigma Y^2) - (\Sigma Y)^2]}}$$

X	Y	X^2	Y^2	XY
10	5	100	25	50
12	6	144	36	72
14	8	196	64	112
16	9	256	81	144
17	10	289	100	170
$\sum X=69$	$\sum Y=38$	$\sum X^2=985$	$\sum Y^2=306$	$\sum XY=548$

$$N\Sigma XY - \Sigma X\Sigma Y = 5(548) - (69)(38) = 2740 - 2622 = 118$$

$$[N\Sigma X^2 - (\Sigma X)^2] = 5(985) - (69)^2 = 4925 - 4761 = 164$$

$$[N\Sigma Y^2 - (\Sigma Y)^2] = 5(306) - (38)^2 = 1530 - 1444 = 86$$

$$r = \frac{N\Sigma XY - \Sigma X\Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2] \times [N\Sigma Y^2 - (\Sigma Y)^2]}} = \frac{118}{\sqrt{164 \times 86}} = \frac{118}{\sqrt{14104}} = \frac{118}{118.76}$$

r = 0.9936 = 0.99

Result:

Karl Pearson's coefficient of correlation (r)= 0.99.

Hence the variables X and Y are highly positively correlated.

Problem 4:

Calculate Pearson Coefficient of Correlation from the following data.

Height of father (in inches)	*Height of son (in inches)*
65	*67*
66	*68*
67	*64*
67	*68*
68	*72*
69	*70*
71	*69*
73	*70*

Solution:

Karl Pearson's coefficient of correlation,

$$r = \frac{[N \times dxdy] - [(\Sigma dx) \times (\Sigma dy)]}{\sqrt{[(N \times \Sigma dx^2) - (\Sigma dx)^2] \times [(N \times \Sigma dy^2) - (\Sigma dy)^2]}}$$

Assumed mean of X= 67 Assumed mean of Y=68

X Series			Y Series			Product of deviation of x and y series
Height of Father	Deviation from assumed mean(67)	Square of deviation	Height of Son	Deviation from assumed mean(68)	Square of deviation	
X	dx=X- 67	dx^2	Y	dy=Y-68	dy^2	dxdy
65	-2	4	67	-1	1	2
66	-1	1	68	0	0	0
67	0	0	64	-4	16	0
67	0	0	68	0	0	0
68	1	1	72	4	16	4
69	2	4	70	2	4	4
71	4	16	69	1	1	4
73	6	36	70	2	4	12
ΣX = 546	Σdx = 10	Σ dx^2 = 62	ΣY = 548	Σdy = 4	Σ dy^2 = 42	Σdxdy = 26

Coefficient of correlation,

$$r = \frac{[N \times dxdy] - [(\Sigma dx) \times (\Sigma dy)]}{\sqrt{[(N \times \Sigma dx^2) - (\Sigma dx)^2] \times [(N \times \Sigma dy^2) - (\Sigma dy)^2]}}$$

Σdxdy = 26, Σdx = 10, Σdy = 4, $\Sigma dx^2 = 62$, $\Sigma dy^2 = 42$, N =8

$$r = \frac{(8 \times 26) - (10 \times 4)}{\sqrt{[8 \times 62 - 10^2] \times [8 \times 42 - 4^2]}} = \frac{208 - 40}{\sqrt{[496 - 100] \times [336 - 16]}} = \frac{168}{\sqrt{396 \times 320}}$$

$$= \frac{168}{\sqrt{126720}} = \frac{168}{355.98} = 0.472$$

Result:

Coefficient of correlation is 0.472. Hence there is positive correlation between the variables.

Spearman's Rank Correlation Coefficient:

Spearman's rank correlation coefficient is defined as

$$\rho = 1 - \frac{6\Sigma d^2}{n[n^2 - 1]}$$

Where d is the difference of two ranks of two variables.

Rank correlation coefficient also lies between -1 and +1.

When the values are repeated,

The rank correlation coefficient is calculated by:

$$\rho = 1 - \frac{6[\Sigma d^2 + C.F]}{n[n^2 - 1]} \qquad C.F = \frac{m(m^2 - 1)}{12} + ...$$

Where m = the number times the value is repeated. C.F.=Correction Factor.

Merits and Demerits of Rank Correlation Coefficient Method:

Merits:

- It is simple to understand and easy to calculate.
- It is very useful in the case of data which are of qualitative nature, like intelligence, honesty, beauty, efficiency etc.
- No other method can be used when ranks are given except this.
- When the actual data are given, this method can also be applied.

Demerits:

- It cannot be used in the case of bi-variate distribution.
- If the number of items is greater than, say 30, the calculation becomes tedious and requires a lot of time.

Case (I): When ranks are given Problem:
Find Spearman's rank correlation coefficient from the following data.

Rank of X	*Rank of Y*
1	*4*
3	*2*
2	*5*
4	*7*
5	*8*
6	*9*
8	*10*

9	*3*
7	*1*
10	*6*

Solution:

R_1	R_2	$d=R_1 - R_2$	d^2
1	4	1- 4 = -3	9
3	2	3 - 2 = 1	1
2	5	2 -5 = -3	9
4	7	4 -7 = =3	9
5	8	5 - 8 = -3	9
6	9	6- 9 = -3	9
8	10	8 - 10 =-2	4
9	3	9 -3=6	36
7	1	7 -1=6	36
10	6	10 -6=4	16
			$\Sigma d^2=138$

Spearman's rank correlation coefficient

$$\rho = 1 - \frac{6\Sigma d^2}{n[n^2 - 1]}$$

Where d= R_1 - R_2, n= Number of pairs of observations.

R_1 = Rank of X , R_2 = Rank of Y

Spearman's rank correlation coefficient

$$\rho = 1 - \frac{6\Sigma d^2}{n[n^2-1]} = 1 - \frac{6(138)}{10(10^2-1)} = 1 - \frac{828}{10(99)} = 1 - \frac{828}{990} = 1 - 0.836$$

$$\rho = 0.164$$

Result:

Spearman's rank correlation coefficient = 0.164

Hence there is low positive correlation between the variables X and Y.

Case (II): When ranks are not given

Find Spearman's rank correlation coefficient from the following data.

X	*Y*
45	*60*
46	*62*
50	*67*
55	*70*
56	*50*
60	*52*
69	*65*
32	*58*
34	*40*

Solution:

X	Y	R_1	R_2	$d=R_1 - R_2$	d^2
45	60	7	5	2	4
46	62	6	4	2	4
50	67	5	2	3	9
55	70	4	1	3	9
56	50	3	8	-5	25
60	52	2	7	-5	25
69	65	1	3	-2	4
32	58	9	6	3	9
34	40	8	9	-1	1
					$\Sigma d^2 = 90$

Spearman's rank correlation coefficient

$$\rho = 1 - \frac{6\Sigma d^2}{n[n^2 - 1]}$$

Where d= R_1 - R_2 , R_1 = Rank of X , R_2 = Rank of Y

n= Number of pairs of observations.

$$\rho = 1 - \frac{6\Sigma d^2}{n[n^2-1]} = 1 - \frac{6(90)}{9(9^2-1)} = 1 - \frac{540}{9(81-1)} = 1 - \frac{540}{9(80)} = 1 - \frac{540}{720}$$

$$\rho = 1 - 0.75 = 0.25$$

Result

Rank correlation coefficient= 0.25

Hence there is positive correlation between X and Y.

Problem:

The ranking of 10 students in two subjects A and B are as follows:

Subject A	6	5	3	10	2	4	9	7	8	1
Subject B	3	8	4	9	1	6	10	7	5	2

Calculate rank correlation coefficient.

Solution: Calculation of Rank Correlation Coefficient

R_1	R_2	$d = (R_1 - R_2)$	$d^2=(R_1 - R_2)^2$
6	3	+3	9
5	8	-3	9
3	4	-1	1
10	9	+1	1
2	1	+1	1
4	6	-2	4
9	10	-1	1
7	7	0	0
8	5	+3	9
1	2	-1	1
		$\Sigma d = 0$	$\Sigma d^2 = 36$

Rank correlation coefficient

$$\rho = 1 - \frac{6\Sigma d^2}{n[n^2 - 1]}$$

$$R, \rho = 1 - \frac{6 \times \Sigma d^2}{N(N^2 - 1)} = 1 - \frac{6 \times 36}{10(10 - 1)} = 1 - \frac{216}{990} = 1 - 0.218 = 0.782$$

Result:

Rank correlation coefficient = 0.782

Hence there is positive correlation between the variables.

Problem:

The competitors in a beauty contest are ranked by three judges in the following order:

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	2	1	6	9
3rd Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty.

Solution:

In order to find out which pair of judges has the nearest approach to common tastes in beauty, we compare Rank correlation coefficient between the judgements of :

i. 1st judge and 2nd judge.

ii. 2nd judge and 3rd judge.

iii. 1st judge and 3rd judge.

R_1	R_2	R_3	$d_1=(R_1-R_2)$	$d_2=(R_2-R_3)$	$d_3=(R_1-R_3)$	$(d_1)^2$	$(d_2)^2$	$(d_3)^2$
1	3	6	-2	-3	-5	4	9	25
6	5	4	+1	+1	+2	1	1	4
5	8	9	-3	-1	-4	9	1	16
10	4	8	+6	-4	+2	36	16	4
3	7	1	-4	+6	+2	16	36	4
2	10	2	-8	+8	0	64	64	0
4	2	3	+2	-1	+1	4	1	1
9	1	10	+8	-9	-1	64	81	1
7	6	5	+1	+1	+2	1	1	4
8	9	7	-1	+2	+1	1	4	1
N=10	N=10	N=10	$\Sigma d_1=0$	$\Sigma d_2=0$	$\Sigma d_3=0$	$\Sigma(d_1)^2=200$	$\Sigma(d_2)^2=214$	$\Sigma(d_3)^2=60$

$$R = 1 - \frac{6 \times \Sigma d^2}{N(N^2-1)}$$

Rank correlation between the judgement of 1st and 2nd judges:

$\Sigma (d_1)^2 = 200$ N=10

$$R_{(I\ and\ II)} = 1 - \frac{6 \times 200}{10(10^2-1)} = 1 - \frac{1200}{990} = 1 - 1.212 = -0.212$$

Rank correlation between the judgement of 2nd and 3rd judges:

$\Sigma (d_2)^2 = 214$ N=10

$$R_{(II\ and\ III)} = 1 - \frac{6 \times 214}{10(10^2-1)} = 1 - \frac{1284}{990} = 1 - 1.297 = -0.297$$

Rank correlation between the judgement of 1st and 3rd judges:

$\Sigma (d_3)^2 = 60$ N=10

$$R_{(I\ and\ III)} = 1 - \frac{6 \times 60}{10(10^2-1)} = 1 - \frac{360}{990} = 1 - 0.364 = +0.636$$

Result

$R_{(\text{I and II})}$ = - 0.212 $R_{(\text{II and III})}$ = - 0.297 $R_{(\text{I and III})}$ = + 0.636

Since the coefficient of correlation is maximum in the judgement of the 1st and 3rd judges, we conclude that they have the nearest approach to common taste in beauty.

Case (iii) When values are repeated (or) Equal ranks

Compute Spearman's rank correlation for the following observation:

Candidate	1	2	3	4	5	6	7	8
Judge X	20	22	28	23	30	30	23	24
Judge Y	28	24	24	25	26	27	32	30

Marks are awarded out of 35.

Solution:

Calculation of Spearman's Rank Correlation

Candidate	Judge X	R_1	Judge Y	R_2	$d=R_1 - R_2$	d^2
1	20	1	28	6	-5	25
2	22	2	24(1)	1.5	+0.5	0.25
3	28	6	24(2)	1.5	+4.5	20.25
4	23(3)	3.5	25	3	+0.5	0.25
5	30(7)	7.5	26	4	+3.5	12.25
6	30(8)	7.5	27	5	+2.5	6.25
7	23(4)	3.5	32	8	-4.5	20.25
8	24	5	30	7	-2	4
N=8						$\Sigma d^2=88.50$

Spearman's rank correlation coefficient

$$\rho = 1 - \frac{6[\Sigma d^2 + C.F]}{n[n^2 - 1]}$$

$$C.F = \frac{m(m^2-1)}{12} + \frac{m(m^2-1)}{12} + \ldots..$$

$$C.F = \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12}$$

$$= \frac{2(4-1)}{12} + \frac{2(4-1)}{12} + \frac{2(4-1)}{12}$$

$$= \frac{2(3)}{12} + \frac{2(3)}{12} + \frac{2(3)}{12} = \frac{6}{12} + \frac{6}{12} + \frac{6}{12} = 0.5 + 0.5 + 0.5 = 1.5$$

Therefore,

$$\rho = 1 - \frac{6[\Sigma d^2 + C.F]}{n[n^2-1]} = 1 - \frac{6[88.5+1.5]}{8(8^2-1)} = 1 - \frac{6(90)}{8(64-1)} = 1 - \frac{540}{8(63)} = 1 - \frac{540}{504}$$
$$= 1 - 1.07 = -0.07$$

Result:

Rank correlation coefficient is -0.07.Hence there is low negative correlation between the variables.

Problem:

Compute Spearman's rank correlation for the following observation:

X	*60*	*58*	*76*	*80*	*64*	*76*	*90*	*92*	*76*	*82*
Y	*40*	*45*	*53*	*45*	*58*	*60*	*59*	*65*	*70*	*62*

Solution

Spearman's rank correlation coefficient

$$\rho = 1 - \frac{6[\Sigma d^2 + C.F]}{n[n^2-1]}$$

$$C.F = \frac{m(m^2-1)}{12} + \frac{m(m^2-1)}{12} + ...$$

UNIT III REGRESSION ANALYSIS

Introduction:

After having established the fact that two variables are closely related, we may be interested in estimating(predicting) the value of one variable given the value of another. For example, if we know that advertising and sales are correlated, we can find out expected amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales. Regression analysis reveals average relationship between two or more variables and this makes possible estimation or prediction.

During the study of hereditary characteristics, Sir Francis Galton found that the height of different groups of sons had the tendency to regress, that is to go back towards the overall average height of all groups of fathers. He called the line of average relationship as the line of regression. It is referred to as the estimating equation because based on that the value of one variable can be found corresponding to a specified value of another variable.

Definition:

"Regression is the measure of the average relationship between two or more variable in terms of the original units of the data. '' - Blair

Uses of Regression Analysis:

1. Regression analysis is used in statistics in all those fields where two or more relative variables are having the tendency to go back to the average.
2. Regression analysis predicts the value of dependent variables from the values of independent variables.
3. With the help of regression coefficients , correlation coefficient can be calculated.
4. In business and economics, it is very helpful to study the predictions.

Difference between Correlation and Regression:

Correlation	*Regression*
1. Correlation is the relationship between two or more variables	*1. Regression is a mathematical measure of average relationship between two or more variables.*
2. It does not indicate the cause and effect relationship between variables.	*2. It indicates the cause and effect relationship between variables.*
3. The coefficient of correlation is a relative measure. The range of relationship lies between ±1.	*3. Regression coefficient is an absolute figure. If we know the value of the independent variable, we can find the value of the dependent variable.*
4. There may be nonsense correlation between two variables.	*4. In regression there is no such nonsense regression.*
5. It has limited applications.	*5. It has wider applications.*
6. It is not very useful for further mathematical treatment.	*6. It is widely used for further mathematical treatment.*

Need for two regression lines:

If we take the case of two variables, X and Y. We shall have two regression lines as the regression of X on Y and the regression of Y on X. The regression line of Y on X gives the most probable values of Y for given values of X and the regression line of X on Y gives the most probable values of X for given values of Y. They should not be interchanged in their usage because as their basic assumptions differ. Hence, there is a need for two regression lines.

Note:

1.If r = ±1, the regression lines will coincide, i.e., we will have only one line.

2.If r = 0, the lines of regression are at right angles.

Regression equation of X on Y is given by

$$(X - \bar{X}) = b_{xy}(Y - \bar{Y})$$

Regression equation of Y on X is given by

$$(Y - \bar{Y}) = b_{yx}(X - \bar{X})$$

Where,

$$b_{xy} = \frac{N\Sigma XY - [(\Sigma X)(\Sigma Y)]}{N\Sigma Y^2 - (\Sigma Y)^2} \quad \text{or} \quad b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$b_{yx} = \frac{N\Sigma XY - [(\Sigma X)(\Sigma Y)]}{N\Sigma X^2 - (\Sigma X)^2} \quad \text{or} \quad b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

Properties of Regression Coefficients:

i. Correlation coefficient is the geometric mean of the two regression coefficients, i.e.,

$$r = \pm\sqrt{b_{xy}b_{yx}}$$

ii. The two regression coefficients and the correlation coefficients have the same sign.

iii. Both the regression coefficients have the same sign.

iv. Both the regression coefficients can not be greater than one numerically simultaneously.

v. Regression coefficients are independent of change of origin but are affected by change of scale.

Note:

1. The two regression lines intersect at .
2. The two regression equations are generally different and are not to be interchanged in their usage.

Problem1:

Find the two regression equations and estimate y when x=40

X	*10*	*12*	*13*	*16*	*17*
Y	*19*	*22*	*24*	*27*	*29*

Solution:

Regression equation of X on Y is $(X - \bar{X}) = b_{xy}(Y - \bar{Y})$

Where $\bar{X} = \frac{\Sigma X}{N} \bar{Y} = \frac{\Sigma Y}{N} b_{xy} = \frac{N \times \Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$

X	Y	XY	X^2	Y^2
10	19	190	100	361
12	22	264	144	484
13	24	312	169	576
16	27	432	256	729
17	29	493	289	841
ΣX=68	ΣY=121	ΣXY=1691	ΣX^2=958	ΣY^2=2991

Calculation of means and regression coefficients

$$\bar{X} = \frac{\Sigma X}{N} = \frac{68}{5} = 13.6 \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{121}{5} = 24.2$$

$$b_{xy} = \frac{N \times \Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)}$$

$$b_{xy} = \frac{5 \times 1691 - (68)(121)}{5 \times 2991 - (121)^2} = \frac{8455 - 8228}{14955 - 14641} = \frac{227}{314} = 0.723$$

$$\begin{aligned} (X - \bar{X}) &= b_{xy}(Y - \bar{Y}) \\ (X - 13.6) &= 0.723(Y - 24.2) \\ X - 13.6 &= 0.723Y - 17.497 \\ X &= 0.723Y - 17.497 + 13.6 \\ X &= 0.723Y - 3.897 \end{aligned}$$

Regression equation of X on Y is $X = 0.723Y - 3.897$.

Regression equation of Y on X is $(Y - \bar{Y}) = b_{yx}(X - \bar{X})$

$\bar{X} = 13.6 \qquad \bar{Y} = 24.2$

$$b_{yx} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

$$b_{yx} = \frac{5 \times 1691 - (68)(121)}{5 \times 958 - (68)^2} = \frac{8455 - 8228}{4790 - 4624} = \frac{227}{166} = 1.367$$

$$\begin{aligned}(Y - \bar{Y}) &= b_{yx}(X - \bar{X}) \\ (Y - 24.2) &= 1.367(X - 13.6) \\ Y - 24.2 &= 1.367X - 18.5912 \\ Y &= 1.367X - 18.5912 + 24.2 \\ Y &= 1.367X + 5.6088\end{aligned}$$

Regression equation of Y on X is $Y = 1.367X + 5.6088$

When X = 40, $Y = 1.367X + 5.6088$

$$\begin{aligned}Y &= 1.367(40) + 5.6088 \\ Y &= 54.68 + 5.6088 \\ Y &= 60.2888 \\ Y &= 60.29\end{aligned}$$

Result:

Regression equation of X on Y is $X = 0.723Y - 3.897$

Regression equation of Y on X is $Y = 1.367X + 5.6088$

And when X is 40, Y is 60.29.

Problem 2:

Find the two regression equations and estimate x when y=50

X	*9*	*10*	*12*	*15*	*18*	*20*
Y	*20*	*18*	*15*	*12*	*16*	*17*

Solution:

Regression equation of X on Y is $(X - \bar{X}) = b_{xy}(Y - \bar{Y})$

$\bar{X} = \frac{\Sigma X}{N}$ $\bar{Y} = \frac{\Sigma Y}{N}$ $b_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$

X	Y	XY	X^2	Y^2
9	20	180	81	400
10	18	180	100	324
12	15	180	144	225
15	12	180	225	144
18	16	288	324	256
20	17	340	400	289
∑X=84	∑Y=98	∑XY=1348	$\sum X^2$=1274	$\sum Y^2$=1638

$\bar{X} = \frac{\Sigma X}{N} = \frac{84}{6} = 14$ $\bar{Y} = \frac{\Sigma Y}{N} = \frac{98}{6} = 16.33$

$$b_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$$

$$b_{xy} = \frac{6 \times 1348 - (84)(98)}{6 \times 1638 - (98)^2} = \frac{8088 - 8232}{9828 - 9604} = \frac{-144}{224} = -0.6429 = -0.64$$

Regression equation of X on Y is

$$\begin{aligned}(X - \bar{X}) &= b_{xy}(Y - \bar{Y}) \\ (X - 14) &= -0.64(Y - 16.33) \\ X - 14 &= -0.64Y - \big(-0.64(16.33)\big) \\ X - 14 &= -0.64Y + 10.45 \\ X &= -0.64Y + 10.45 + 14 \\ X &= -0.64Y + 24.45\end{aligned}$$

Regression of Y on X is $(Y - \bar{Y}) = b_{yx}(X - \bar{X})$

$\bar{X} = 14$ $\bar{Y} = 16.33$

$$b_{yx} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

$$b_{yx} = \frac{6 \times 1348 - (84)(98)}{6 \times 1274 - (84)^2} = \frac{8088 - 8232}{7644 - 7056} = \frac{-144}{588} = -0.24$$

$$(Y - \bar{Y}) = b_{yx}(X - \bar{X})$$

$$(Y - 16.33) = -0.24(X - 14)$$
$$Y - 16.33 = -0.24X - \big(-0.24(14)\big)$$
$$Y - 16.33 = -0.24X + 3.36$$
$$Y = -0.24X + 3.36 + 16.33$$
$$Y = -0.24X + 19.69$$

Regression equation of Y on X is $Y = -0.24X + 19.69$.

When Y=50, $X = -0.64Y + 24.45$

$$X = -0.64(50) + 24.45$$
$$X = -32 + 24.45$$
$$X = -7.55$$

Result:

Regression equation of X on Y is $X = -0.64Y + 24.45$.

Regression equation of Y on X is $Y = -0.24X + 19.69$.

When value of Y is 50, value of X is -7.55.

Problem 3:

The following data relate to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales in thousand rupees:

Salesmen	A	B	C	D	E	F	G	H	I
Intelligence	50	60	50	60	80	50	80	40	70
Weekly Sales	30	60	40	50	60	30	70	50	60

a) Obtain the regression equations of sales on intelligence test scores of the salesmen and intelligence test score on weekly sales

b) If the intelligence test score of a salesman is 65, what would be his expected weekly sales?

Solution:

Let intelligence test score be X and weekly sales be Y.

X	$(X - \bar{X})$	$(X - \bar{X})^2$	Y	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
50	-10	100	30	-20	400	200
60	0	0	60	+10	100	0
50	-10	100	40	-10	100	100
60	0	0	50	0	0	0
80	+20	400	60	+10	100	200
50	-10	100	30	-20	400	200
80	+20	400	70	+20	400	400
40	-20	400	50	0	0	0
70	+10	100	60	+10	100	100
$\Sigma X = 540$	$\Sigma (X - \bar{X}) = 0$	$\Sigma (X - \bar{X})^2 = 1600$	$\Sigma Y = 450$	$\Sigma (Y - \bar{Y}) = 0$	$\Sigma (Y - \bar{Y})^2 = 1600$	$\Sigma (X - \bar{X})(Y - \bar{Y}) = 1200$

$\bar{X} = \frac{\Sigma X}{N} = \frac{540}{9} = 60 \bar{Y} = \frac{\Sigma Y}{N} = \frac{450}{9} = 50 \qquad b_{yx} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\Sigma (X - \bar{X})^2} = \frac{1200}{1600} = 0.75$

$$b_{xy} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\Sigma (Y - \bar{Y})^2} = \frac{1200}{1600} = 0.75$$

Regression equation of Y on X is

$$\begin{aligned} (Y - \bar{Y}) &= b_{yx}(X - \bar{X}) \\ (Y - 50) &= 0.75(X - 60) \\ Y - 50 &= 0.75X - (0.75 \times 60) \\ Y - 50 &= 0.75X - 45 \\ Y &= 0.75X - 45 + 50 \\ Y &= 0.75X + 5 \end{aligned}$$

Regression equation of X on Y is

$$\begin{aligned} (X - \bar{X}) &= b_{xy}(Y - \bar{Y}) \\ (X - 60) &= 0.75(Y - 50) \\ X - 60 &= 0.75Y - (0.75 \times 50) \\ X - 60 &= 0.75Y - 37.5 \\ X &= 0.75Y - 37.5 + 60 \\ X &= 0.75Y + 22.5 \end{aligned}$$

Expected weekly sales when intelligence test score of a salesman is 65.

$$Y = 0.75X + 5$$

Putting X = 65, $Y = 0.75(65) + 5$

$$Y = 48.75 + 5$$

$$Y = 53.75$$

Result:

Regression equation of X on Y is$X = 0.75Y + 22.5$.

Regression equation of Y on X is $Y = 0.75X + 5$.

When intelligence test score of a salesman is 65, his expected sales is 53.75.

Problem 4:

You are given the following data:

	X	Y
Arithmetic Mean	36	85
Standard Deviation	11	8

Correlation coefficient between X and Y is 0.66

i) Find the two regression equations.

ii) Estimate the value of X when Y= 75.

Solution:

i) Regression equation of X on Y is $(X - \bar{X}) = b_{xy}(Y - \bar{Y})$

$\bar{X} = 36 \qquad \bar{Y} = 85 \qquad \sigma_x = 11 \qquad \sigma_y = 8 \, b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$$(X - 36) = 0.66\frac{11}{8}(Y - 85)$$

$$(X - 36) = 0.9075(Y - 85)$$

$$X - 36 = 0.9075Y - 77.1375$$

$$X = 0.9075Y - 77.1375 + 36$$

$$X = 0.9075Y - 41.1375$$

ii) Regression equation of Y on X is $(Y - \bar{Y}) = b_{yx}(X - \bar{X})$

$\bar{X} = 36 \qquad \bar{Y} = 85 \qquad \sigma_x = 11 \qquad \sigma_y = 8 \, b_{yx} = r\frac{\sigma_y}{\sigma_x}$

$$(Y - 85) = 0.66\frac{8}{11}(X - 36)$$

$$(Y - 85) = 0.48(X - 36)$$

$$Y - 85 = 0.48X - 17.28$$

$$Y = 0.48X + 67.72$$

ii. From the regression equation of X on Y, we can find out the estimated value of X when Y =75;

$$X = 0.9075Y - 41.1375$$

$$= 0.9075(75) - 41.1375$$

$$= 68.0625 - 41.137$$

$$X = 26.925$$

Result:

i. Regression equation of X on Y is$X = 0.9075Y - 41.1375$

.

Regression equation of Y on X is$Y = 0.48X + 67.72$

.

ii. When Y=75, the value of X is 26.925

UNIT IV SAMPLING METHODS

Population and Sample:

Aggregate of all the units of an investigation is called population. For example: If the investigation is on the housing conditions of a locality, all the houses in that locality constitute the population.

A Sample is a part of a population or few units of a population.

For example: A handful of rice taken from a sack of rice constitute the sample.

Methods to collect data:

1.Census Methods

2.Sample Method

Census Method or Census survey:

Under Census method, the information is collected from each and every unit of the population. For example, if we study the average expenditure of students of MK University and if there are 5000 students studying in that university, we must study the expenditure of all 5000 students. This method is known as Census method.

Merits:

1. The data are collected from each and every unit of the population.
2. The results are more accurate and reliable.
3. Intensive study is possible.
4. The data collected may be used for various surveys etc.

Demerits:

1. It requires more time, money, energy etc.
2. It is a costly method.
3. The government only can use this method.
4. It is not possible where the universe is infinite.

Sample Method or Sample Survey or Sample Enquiry

Only a part of the population will be studied in the case of sample survey.

For e.g., A housewife tests a small quantity if rice to see whether it has been well cooked but will not inspect all the rice.

Merits:

1. It saves time, money and energy.
2. Sampling provides more detailed information about an item.
3. More reliable research can be obtained.
4. Sampling method is sometimes the only method possible if the population is infinite.
5. Administrative convenience.
6. More scientific.

Demerits:

1. Choosing a representative sample is difficult.
2. Only experts can properly generalize the result of sample to the population.
3. If the sample is not representative of the population, the result may be false and misleading.
4. There may be personal biases and prestige regarding the choice of technique and drawing of sampling unit.
5. Inaccurate result may be obtained if the size of the sample is inadequate.

Methods of Sampling

1. Random Sampling methods or Probability Sampling methods.
    a. Simple or Unrestricted Random Sampling.
    b. Restricted Random Sampling
        i. Systematic Random Sampling.
        ii. Stratified Random Sampling.
        iii. Cluster Random Sampling or multi-stage sampling.
2. Non-Random Sampling methods.
    a. Judgement Sampling.
    b. Convenience Sampling.
    c. Quota Sampling.

Simple Random Sampling

Simple Random Sampling is a sampling technique in which each and every unit of the population has an equal opportunity of being selected in the sample. There are two methods of random selection:

a. Lottery Method
b. Table of random numbers

Lottery Method:

The procedure can be easily described by considering an example. Let a sample of size 25 be required from the population of 200. Prepare 200 cards or lots (pieces of paper) of same size and color. Write on each of them the name or other distinguishing mark of one unit of the population. Fold them uniformly and shuffle them well. 25 of them are to be selected.

Table of Random Methods:

The units of the population are first assigned numbers for identification. Consecutive numbers are noted from the table of random numbers. Those units of population constitute the sample.

For example, let a random sample of 30 students be required from a college where 914 students are studying. We assign the students three-digit numbers 000,001,002….913. We start at any 3-digit number and make note of the subsequent numbers row wise or column wise.

Merits and Demerits of Simple Random Sampling:

Merits:

1. Personal bias is eliminated.
2. Compared with non-random sampling, any random sampling is better.
3. There is no need for the thorough knowledge about the members of the population.
4. The accuracy of sample can be tested by examining another sample from the same universe.

Demerits:

1. Preparing cards or making use of random number tables is tedious.
2. Where there are large differences between the population, stratified random sampling is better than simple random sampling.

3. A later sample is needed by this method than stratified random sampling method for getting reliable estimate.

Systematic Sampling:

One member is selected from the first k members of the population by simple random sampling method and every k^{th} member joins the sample afterwards.k is known as sampling interval.

If a sample size of 25 is required from a population of 200:

$$k = \frac{Size\ of\ population}{Size\ of\ sample} = \frac{200}{25} = 8$$

If a member, say 7, is selected from the first 8 members by lottery method, the other members of the sample are (7+8)15, (15+8)23 , (23+8)31…. 199.

Merits and Demerits of Systematic Sampling

Merits:

1. It is more convenient and easier method than simple random sampling.
2. Time and work involved are less.
3. It is more efficient, and results obtained are found to be more satisfactory.

Demerits:

The greatest limitation is in finding whether there is any periodic variation in the population and if so, in finding whether that period coincides with the sampling interval.

Stratified Sampling:

Under this method, the population is divided into different groups or classes called strata and a sample is drawn from each stratum at random. The aggregate of samples of all strata is called stratified random sample.

For example, if we are interested in a study of the unemployment problem in India, we divide it into a number of regions and from each region a sample is made. Aggregate of all samples from the regions is stratified random sample. There are two types of Stratified Random Sampling:

a. Proportional Stratified Sampling.
b. Disproportional Stratified Sampling.

Proportional Stratified Sampling:

In Proportional Stratified Sampling plan, the number of items drawn from each stratum is proportional to size of the stratum. For e.g. if the population is divided into five groups, the respective size being 10, 15, 20, 30, 25 percent of the population and a sample of 500 drawn, the desired proportional sample may be obtained in the following manner:

From the first stratum	500 x 0.10	=	50
From the Second stratum	500 x 0.15	=	75
From the Third stratum	500 x 0.20	=	100
From the Fourth stratum	500 x 0.30	=	150
From the Fifth stratum	500 x 0.25	=	125
	Total		500

Disproportional Stratified Sampling:

In disproportional stratified sampling, an equal number of items is taken from each stratum regardless of size of the stratum.

Merits and Demerits of Stratified Sampling

Merits:

1. It is simple to understand.
2. This is an accurate method.
3. This method covers a greater geographical area than other methods.
4. It is easy to administer as the universe is subdivided.
5. It is more representative method than other methods

Demerits:

1. This is a time consuming and expensive method.
2. Maximum care must be exercised in dividing the population into various strata.
3. This method is suitable only when the population is heterogeneous.

Cluster Sampling

It refers to a sampling procedure, which is carried out in several stages. The whole population is divided into sampling units, and these units are again divided into sub-units. This process will continue till we reach a least number.

For example, we want to take 5000 students from Tamil Nadu. We must take universities at the stage, then the number of colleges at the second stage, selection of students at the third stage etc.

Merits:

1. It introduces flexibility in the sampling method.
2. It is helpful in large-scale survey.
3. It is valuable in underdeveloped countries, where no detailed and accurate framework is available.

Demerits:

1. It is less accurate than other methods.

Judgement Sampling

Under this method, the investigator selects those units which he considers to be typical. He believes that those units possess the characteristics of the population.

Merits:

1. It is an easy and popular method.
2. It can accommodate few units which are to be obliged.
3. It saves time and cost.

Demerits:

1. It is not scientific.
2. The size of the sample could not be determined.
3. The magnitude of the sampling error cannot be calculated.
4. It is sometimes very biased.

Convenience Sampling

The sample units are selected according to the convenience of the investigator. A sample obtained from readily available lists such as telephone directories is of this kind. The term **chunk** is used in this context. **Chunk** refers to a sample which is selected neither at random nor by judgment but by convenience.

Merits:

1. It is easy, cheap and time saving.
2. It is useful in pilot studies.

Demerits:

1. It does not provide a representative sample.
2. It is biased.
3. It is not a popular method.

Quota Sampling

Number of units to be selected from each sub-group of the population is decided according to certain criteria. Each interviewer is asked to meet a fixed number of units. These units are his **quota**.

Merits:

1. The cost per unit selected is considerably less.
2. It is popular in market surveys and opinion polls.

Demerits:

1. Quota sampling is neither scientific nor expected to yield good results.
2. The interviewer may not choose typical units because of their ignorance of the population.

Sampling Error:

There will be difference between the actual population value and its value estimated from the sample. It is called sampling error.

Non-Sampling Error:

Errors which occur in both the sample survey and census survey is called Non-sampling error.

Basic Principles of Sampling:

1. Law of Statistical Regularity.
2. Law of Large numbers.

Law of Statistical Regularity:

The Law of Statistical Regularity lays down that a moderately large group are almost sure on the average to possess the characteristics of the population.

Law of Large Numbers:

According to this law, the larger is the size of the sample then greater is the amount of accuracy.

Essentials of Sampling:

1. The sample must be representative.
2. It must be homogeneous.
3. Size of the sample must be adequate.
4. Optimization. The effect must be to get maximum result both in terms of cost as well as efficiency.

UUIT –V
TESTING OF HYPOTHESIS

Introduction

Hypothesis means a statement about population Parameter,

Parameter

Statistical constants of the population viz., mean(μ) , variance(σ^2) etc., which are usually referred to as parameter.

Statistic

Statistical measures computed from the sample observations alone, e.g., mean($\bar{x}$), variance(s^2) etc., have been termed as statistic.

Sampling distribution of a statistic

The set of the values of the statistic so obtained , one for each sample, constitutes what is called the sampling distribution of the statistic.

Standard error

The standard deviation of the sampling distribution of a statistic is known as its standard error and abbreviated as S.E

S.No.	*Statistic*	*Standard Error*
1	*Sample mean :* $\bar{x}$	$\sigma/\sqrt{n}$
2	*Observed sample proportion: p*	$\sqrt{PQ/n}$
3	*Sample standard deviation :s*	$\sqrt{\sigma^2/2n}$
4	*Difference of two sample means:* $(\overline{x_1} - \overline{x_2})$	$\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$
5	*Difference of two sample standard deviations:* $(s_1 - s_2)$	$\sqrt{\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}}$
6	*Difference of two sample proportions:* $(p_1 - p_2)$	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

Utility of standard error

Standard Error(S.E) plays a very important role in the larger sample theory and forms the basis of the testing of hypothesis.

(i) The magnitude of the standard gives an index of the precision of the estimate of the parameter. The reciprocal of the standard error is taken as a measure of reliability of precision of the statistic.

S.E(p) = $\sqrt{PQ/n}$ and

S.E($\bar{x}$) = $\sigma/\sqrt{n}$

(ii) S.E enable us to determine the probable limits* within which the population parameter may be expected to lie. For example, the probable limits for population proportion P are given by

$$p \pm 3\sqrt{pq/n}$$

Tests of significance

The study of the tests of significance, which enable us to decide on the basis of the sample results,if

(i) the deviation between the observed sample statistic and the hypothetical parameter value or,

(ii) the deviation between two independent sample statistic is significant or might be attributed to chance or the fluctuations of sampling.

Null hypothesis

Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true, usually denoted by H_0 i,e., hypothesis of no difference is called null hypothesis.

For example ,in case of a single statistic, H_0 will be the sample statistic does not differ significantly from the hypothetical parameter value. i.e., H_0: $\mu = \mu_0$ **Alternative hypothesis**

Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis, usually denoted by H_1.

For example, if we want to test the null hypothesis H_0: $\mu = \mu_0$ then the alternative hypothesis could be:

(i) H_1: $\mu \neq \mu_0$ (i.e., $\mu > \mu_0$ or $\mu < \mu_0$)
(ii) H_1 : $\mu > \mu_0$
(iii) H_1: $\mu < \mu_0$

Errors in sampling

Type I Error

The error of rejecting the null hypothesis (accepting H_1) when H_0 is true is called Type I Error.

Type II Error

The error of accepting the null hypothesis when H_0 is false (H_1 is true) is called Type II Error.

Note

Type 1 error amounts to rejecting a lot when it is good and Type II Error may be regarded as accepting the lot when it is bad.

Thus P{ Reject a lot when it is good} = α

and P{ Accept a lot when it is bad} = β

Where α and β are referred to as producer's risk and consumer's risk respectively.

Level of of significance

The probability of Type 1 error is known as level of significance. it is denoted by α

Critical region

A region corresponding to a statistic t in the sample space S which amounts to rejection of H_0 is termed as critical region or rejection region of the statistic.

Critical values or significant values

The value of test statistic which separates the critical region(or rejection) region and the acceptance region is called the critical value or significant value.

One tailed and Two tailed tests

In any test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic.

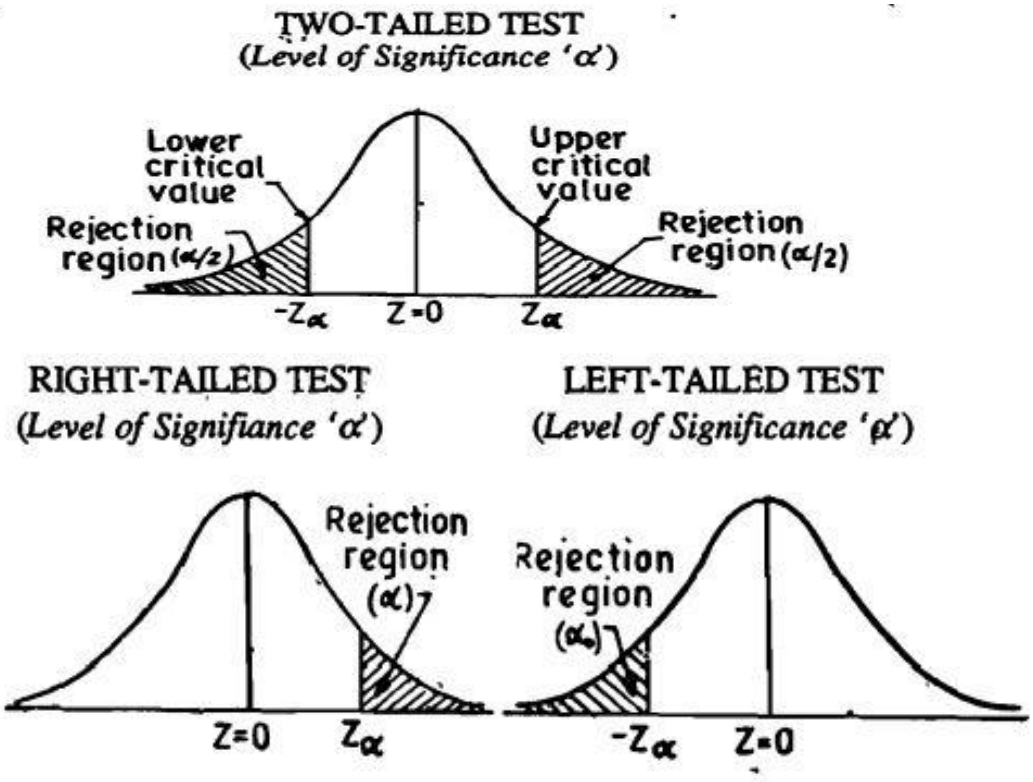
A test of any statistical hypothesis that the alternative hypothesis is one tailed (right-tailed or left- tailed) is called a one tailed test. For example a test for testing the mean of a population H_0: $\mu = \mu_0$ against the alternative hypothesis:

$H_1 : \mu > \mu_0$ (right- tailed test)

H_1: $\mu < \mu_0$ (left- tailed test), is a single tailed test.

In the right tailed test , the critical region lies entirely in the right tail of the distribution, while for the left –tailed test H_1: $\mu < \mu_0$ critical region lies entirely in the left tail of the distribution.

A test of statistical hypothesis where the alternative hypothesis is two tailed such as H_1: $\mu \neq \mu_0$ (i.e., $\mu > \mu_0$ or $\mu < \mu_0$) is known as two tailed test and in such a case the critical region is given by the portion of the area lying in both tails of the probability curve of the statistic.



Procedure for testing of hypothesis

we now summarise below the various steps in testing of a statistical hypothesis in a systematic manner.

1. Null hypothesis: Set up the null hypothesis H_0
2. Alternative hypothesis: Set up the alternative hypothesis H_1. This will enable us to decide whether we have to use a single tailed (right of left) test or two tailed test.
3. Level of significance: Choose the appropriate level of significance(α) depending on the reliability of the estimates and permissible risk. This is to be decided before the sample is drawn, i.e., α is fixed in advance.
4. Test statistic (or test criterion)

   Compute the test statistic

   $Z = \frac{t-E(t)}{S.E(t)}$, under H_0

5. Conclusion : We compare the computed value of Z with tabulated value at the given level of significance α.

If $|Z|$ < tabulated value, null hypothesis may be accepted at given level of significance.

If $|Z|$ > tabulated value, null hypothesis may be rejected at given level of significance.

Test of significance for single proportion

Null hypothesis H_0

There is no significant difference between the sample proportion and population proportion .

Alternative hypothesis H_1

There is a significant difference between the sample proportion and population proportion.

Level of significance α

Either 5% or 1%

Test Statistic:

Under the null hypothesis, the test statistic is

$$Z = \frac{p-P}{S.E(p)} \sim N(0,1)$$

$$= \frac{p-P}{\sqrt{PQ/n}} \sim N(0,1)$$

Conclusion:

(i) If the calculated value of $|Z|$ is less than the tabulated value, the null hypothesis may be accepted.

(ii) If the calculated value of $|Z|$ is greater than the tabulated value, the null hypothesis may be rejected.