BALAJI INSTITUTE OF I.T AND MANAGEMENT KADAPA

STATISTICS FOR MANAGERS

(21E00105)

ICET CODE: BIMK

www.bimkadapa.in

1ST & 2ND INTERNAL EXAM

Name of the Faculty: T.HIMMAT

Units covered: 1-5 UNITS

E-Mail: himmatbimk@gmail.com

MASTER OF BUSINESS ADMINISTRATION
MBA; MBA (General Management); MBA (Business Management)
COMMON COURSE STRUCTURE

Course Code	**STATISTICS FOR MANAGERS**	**L**	**T**	**P**	**C**
21E00105		**4**	**0**	**0**	**4**
	Semester	**I**			

Course Objectives:

- To explain descriptive statistics and inferential statistics
- To introduce various measurements used to describe the data and inter the results of the data analysis.
- To describe the concept of probability, theorems, and types of probability distributions of data.
- To impart the computational, analytical and interpretation skills using the data

Course Outcomes (CO): Student will be able to

- Understand statistical techniques popularly used to describe the data in managerial decision making.
- Know the procedure involved in inferential statistics and appropriate tests for given data.
- Learn the computational skill , interpretation of results of the data analysis.
- Analyse and differentiate various types of data distribution and its probability distribution.

UNIT - I Lecture Hrs: 12

Introduction of statistics – Nature & Significance of Statistics to Business, , Measures of Central Tendency: Mean – Median – Mode ; Measures of Dispersion: range, quartile deviation, mean deviation, standard deviation, coefficient of variation.

UNIT - II Lecture Hrs: 12

Correlation & Regression : Introduction, Significance and types of correlation – Measures of correlation – Co-efficient of correlation. Regression analysis – Meaning and utility of regression analysis – Comparison between correlation and regression – Properties of regression coefficients-Rank Correlation.

UNIT - III Lecture Hrs:12

Probability – Meaning and definition of probability – Significance of probability in business application – Theory of probability: Addition and multiplication – Binominal distribution– Poisson distribution – Normal distribution.

UNIT - IV Lecture Hrs:12

Testing of Hypothesis- Hypothesis testing: One sample and Two sample tests for means and proportions of large samples (z-test), One sample and Two sample tests for means of small samples (t-test), ANOVA Test : One-way and two way ANOVA .

UNIT - V Lecture Hrs: 08

Non-Parametric Methods: Importance of Non-Parametric method – difference between parametric and non-parametric methods**;** Chi-square test : Test of Goodness of fit - test for Independence of Attributes; Sign test: One sample and paired samples data.

Textbooks:

1. Statistical Methods, Gupta S.P., S.Chand.Publications
2. Business Statistics, J.K.Sharma, Vikas house publications house Pvt Ltd

Reference Books:

1. Statistics for Management, Richard I Levin, David S.Rubin, Pearson,
2. Complete Business Statistics, Amir D. Aezel, Jayavel, TMH,
3. Statistics for Management, P.N.Arora, S.Arora, S.Chand
4. Statistics for Management ,Lerin, Pearson Company, New Delhi.
5. Business Statistics for Contemporary decision making, Black Ken, New age publishers.
6. Business Statistics, Gupta S.C & Indra Gupta, Himalaya Publishing House, Mumbai

Online Learning Resources:
https://onlinecourses.swayam2.ac.in/cec20_mg13/preview https://onlinecourses.nptel.ac.in/noc20_mg23/preview https://iimbx.iimb.ac.in/statistics-for-business-i/

UNIT-1

Introduction To Statistics

Meaning of Statistics :-

Statistics means collecting the data, organizing the data, presenting the data, analysing the data and interpreting the data.

Meaning of statistics in 2 Senses. Singular Sense and plural sense.

In Singular Sense :-

In Singular sense, statistics means collection, organizing, presenting, analyzing and interpreting the data.

In Plural Sense :-

* In plural sense, statistics means collection of numerical facts.

* In this sense not only consider figures but also take percentages, averages and co-efficient derived from numerical facts.

Nature and Significance of Statistics to Business :-

Statistics is used, to extended the different fields of experiment to draw valid conclusions, and it is used to found the increase importance and usage. The nature and importance of statistics in various fields are listed given below.

* State Affairs :-

- * To collect the information and study the economic condition of people in the states.
- * To assess the resources available in states.
- * To help state to take decision on accepting (or) rejecting it's policy based on statistics.
- * To provide information and analysis on various factors of state like wealth, crimes, agriculture experts, education etc.,

* Economics :-

- * Helps in formulation of economic laws and policies.
- * Helps in studying economic problems.
- * Helps in compiling the national income accounts.
- * Helps in economic planning.

* Business :-

- * Helps to take decisions on location and size.
- * Helps to study demand and supply.
- * Helps in forecasting and planning.
- * Helps controlling the quality of the products (or) process.
- * Helps in making marketing decisions.
- * Helps for production, planning and inventory management.
- * Helps in business risk analysis.

Subject : Date :
Title of the test case :
Case study No. : Page No. :

* Education :-

statistics is necessary to formulate the policies regarding start of new courses, consideration of facilities available for proposed courses.

* Accounts And Audits :-

* Helps to study the correlation between profits and dividends enable to know trend of future profits.

* In auditing sampling techniques are followed.

* Measure of Central Tendency :-

According to Simpson and Kafka, "Measure of central Tendency is a single value with in the range of the entire mass of data that is used to represent the whole data".

Characteristics :-

* It should be strongly defined :-

An average should be strongly defined so that there is no confusion in regard to it's meaning. If it is not well defined it may be influenced by the prejudice value to represent the distribution.

* It's definition should be in the form of a Mathematical formula :-

With mathematical formulation different persons may not interpret it differently and anybody computing the average form a set of data.

* It should be easy to calculate & Simple to follow :-

An average should be simple in comprehension So that it can be calculated with reasonable case and it's use will be very limited.

* It should be based on all observations in the series :-

An average will be truly representative of the whole mass of data if it is computed from all the observations.

* It should be capable of further algebraic treatment :-

An average should extend itself readly to further algebraic treatment.

* It should be capable of being used in further statistical compitation of processing :-

An average should possess this quality.

For Eg :- * Arithmetic mean is suitable for calculating standard deviation else, it will not be of great use in further statistical analysis and it's utility will be limited.

Subject :		Date :	
Title of the test case :			
Case study No. :		Page No. :	

*** It should possess Sampling stability :-**

An average should be least affected by Sampling. By this we mean that if we take independent random samples of the same size from a given population.

Types Of Averages :-

The following are the important types of averages.

- * Arithmetic mean.
- * Median.
- * Mode.
- * Geometric Mean.
- * Hormonic Mean.

Arithmetic Mean :-

Simple Arithmetic Mean :-

In Individual Series, the process of computing arithmetic mean is the ratio between sum of the observations to the total number of observations. i.e., Symbolically, it is denoted by

Individual Series by direct Method :-

Arithmetic mean $\boxed{A.M., \bar{x} = \frac{\Sigma x}{N}}$.

Where, $\Sigma x = x_1 + x_2 + x_3 + \cdots\cdots + x_n$.

N = Total Observations.

Individual Series by shortcut Method :-

$$\boxed{A.M., \bar{x} = A + \frac{\Sigma d}{N}}$$

Where, A = Assumed mean.

$d = x - A$

(difference between observation to the assumed mean)

Subject	:	Date	:
Title of the test case	:		
Case study No.	:	Page No.	:

Problems :-

Individual Series :-

1) The monthly income of 5 persons are given below. find arithmetic mean.

132, 140, 144, 136 and 148.

Sol:- Direct Method :-

Given 5 persons monthly income are 132, 140, 144, 136, 148.

x
132
140
144
136
148
$\Sigma x = 700$

Where, $\Sigma x = 700$

$N = 5$

$$A.M., \bar{x} = \frac{\Sigma x}{N} = \frac{132+140+144+136+148}{5}$$

$$= \frac{700}{5}$$

$$= 140.$$

$$\boxed{\therefore \bar{x} = 140.}$$

Short-cut Method :-

Where, Assumed mean A = 140.

x	$d = (x-A)$
132	-8
140	0
144	4
136	-4
148	8
	$\Sigma d = 0$

$$A.M., \bar{x} = A + \frac{\Sigma d}{N}$$

$$= 140 + \frac{0}{5}$$

$$= 140 + 0$$

$$= 140.$$

$$\boxed{\therefore \bar{x} = 140.}$$

Arithmetic in Discrete Series :-

Direct Method :-

(i) Multiply each item/variable to it's frequency i.e., $f \times x$.

(ii) Add all the fx values i.e., Σfx.

(iii) Add sum of all the frequencies i.e., $N = \Sigma f$.

Symbolically, it is denoted by

$$\boxed{A.M., \bar{x} = \frac{\Sigma fx}{\Sigma f.}}$$

Shortcut Method :-

(i) Assume any one of the variable in the given series i.e., A for easy calculations.

(ii) Find the deviation value i.e., $d = x - A$ (difference between variable to the assumed mean)

(iii) Multiply the frequencies with all the deviation values

Subject :   Date :

Title of the test case :

Case study No. :   Page No. :

i.e., fd. Then add all the values i.e., Σfd.

Sum all the frequencies i.e., Σf.

Symbolically, it is denoted by.

$$\boxed{A.M., \bar{x} = A + \frac{\Sigma fd}{N}}$$

Where, A = Assumed mean.

$d = x - A$

$N = \Sigma f$ (sum of all frequencies).

Problem :-

1) calculate Arithmetic mean from the following data.

Marks (x) :	20	30	40	50	60	70
No. of students (f) :	8	12	20	10	6	4

Sol:- Direct Method :-

Marks (x)	No. of Students (f)	fx
20	8	160
30	12	360
40	20	800
50	10	500
60	6	360
70	4	280
	N=60	$\Sigma fx = 2460$

$$A.M. \bar{x} = \frac{\Sigma fx}{\Sigma f \text{ or } N}$$

$$= \frac{2460}{60}$$

$$\bar{x} = 41.$$

$$\boxed{\therefore A.M, \bar{x} = 41.}$$

Short cut Method :-

Here, $A = 20$

Marks(x)	Frequency(f)	$d = x - A$	fd
20	8	0	0
30	12	10	120
40	20	20	400
50	10	30	300
60	6	40	240
70	4	50	200
	$N = 60$		$\Sigma fd = 1260$

$$A.M.\ \bar{x} = A + \frac{\Sigma fd}{N}$$

$$= 20 + \frac{1260}{6}$$

$$= 20 + 21$$

$$\bar{x} = 41$$

$$\boxed{\therefore A.M, \bar{x} = 41.}$$

Arithmetic Mean for Continuous Series :-

Direct Method :-

$$\boxed{A.M.\ \bar{x} = \frac{\Sigma fm}{\Sigma f}}$$

(i) Find mid value of each class interval. i.e., 'm'.

$$\text{midvalue} = \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

(ii) Multiply each midvalue of the class by its frequency i.e., fm.

(iii) Sum all the multiply of each midvalue and frequency i.e., Σfm.

(iv) Sum all the frequencies of the class intervals. i.e., Σf (or) N.

Shortcut Method :-

(i) Find midvalue of each class interval i.e., 'm'.

$$\text{Midvalue} = \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

(ii) Assume one value in the midvalues sequency i.e., 'A'.

(iii) Calculate each midvalue deviation i.e., $d = m - A$

Multiply each deviation value with it's frequency, i.e, fd

~~M~~ Sum all the fd values i.e., Σfd.

Sum all the frequencies i.e., Σf

$$\boxed{A.M,\ \bar{x} = A + \frac{\Sigma fd}{\Sigma f}}$$ where, $d = x - A$

Σf = sum of the frequencies

(or)

$$\boxed{\bar{x} = A + \frac{\Sigma fd}{N}}$$

where, $d = x - A$

N= Sum of all the frequencies.

1) Calculate Arithmetic mean from the following data.

Marks :-	0-10	10-20	20-30	30-40	40-50	50-60
Frequency :-	5	10	25	30	20	10

Sol:- Direct Method :-

Marks (x)	Frequency (f)	Midvalues (m)	fm
0-10	5	$\frac{0+10}{2} = 5$	25
10-20	10	$\frac{10+20}{2} = 15$	150
20-30	25	$\frac{20+30}{2} = 25$	625
30-40	30	$\frac{30+40}{2} = 35$	1050
40-50	20	$\frac{40+50}{2} = 45$	900
50-60	10	$\frac{50+60}{2} = 55$	550
	$\Sigma f = 100$		$\Sigma fm = 3,300$

$$A.M, \bar{x} = \frac{\Sigma fm}{\Sigma f}$$

$$= \frac{3,300}{100}$$

$$\bar{x} = 33.$$

$$\boxed{\therefore A.M, \bar{x} = 33.}$$

Short cut Method :-

Here

$A = 5.$

Subject :

Date :

Title of the test case :

Case study No. :

Page No. :

x	f	M	d = m-A	fd
0-10	5	5	0	0
10-20	10	15	10	100
20-30	25	25	20	500
30-40	30	35	30	900
40-50	20	45	40	800
50-60	10	55	50	500
	$\Sigma f = 100$			$\Sigma fd = 2,800$.

$$A.M, \bar{x} = A + \frac{\Sigma fd}{N}$$

$$= 5 + \frac{2800}{100}$$

$$= 5 + 28$$

$$\bar{x} = 33.$$

$$\boxed{\therefore A.M, \bar{x} = 33.}$$

Weighted Arithmetic Mean :-

Arithmetic mean gives equal importance of all the items. But in the case of weighted arithmetic mean the relative importance of different items is not the same for this we compute the weighted arithmetic mean.

The term weighted stands for the relative importance of different items.

$$W.A.M. \boxed{\bar{x}_w = \frac{\Sigma wx}{\Sigma w}.}$$

where wx = product of both variable and their relative weights.

1) The mean height of 25 male workers in a factory is 61 inchs and the mean height of 35 female workers in the same factory is 58 inchs. Find weighted Arithmetic mean of 60 workers in the factory?

A) Male workers in a factory, $w_1 = 25$

Male workers height, $x_1 = 61$ inchs.

Female workers in a factory, $w_2 = 35$

Female workers height, $x_2 = 58$ inchs.

$$\text{W.A.M}, \bar{x}_{w_1 \cdot w_2} = \frac{w_1x_1 + w_2x_2}{w_1 + w_2}$$

$$= \frac{25\times61 + 35\times58}{25+35}$$

$$= \frac{1525 + 2030}{60}$$

$$= \frac{3555}{60}$$

$$= 59.25$$

$$\boxed{\therefore \bar{x}_{w_1w_2} = 59.25}$$

Geometric Mean :-

Geometric mean is defined as the n^{th} root of the product of 'n' items (or) values. If there are 2 items we take the square root. there are 3 items cuberoot is taken and so.........on.

Subject : Date :

Title of the test case :

Case study No. : Page No. :

$$\therefore G.M = \sqrt[n]{x_1 \times x_2 \times x_3 \times \dots \times x_n}$$

When the number of item is more than '3' the task of multiplying the numbers and exactly root because become more difficult to calculate. For this simple calculations logarithms are use.

In Individual Series :- $G.M = A.L\left[\frac{\Sigma \log x}{N}\right]$

In Discrete Series :- $G.M = A.L\left[\frac{\Sigma f \log x}{\Sigma f}\right]$

In Continuous Series :- $G.M = A.L\left[\frac{\Sigma f \log m}{\Sigma f}\right]$

Note :- Calculating of Anti logarithms i.e., A.L = Shift + log + value in scientific calculator.

Problems :-

Individual Series :-

1) Calculate Geometric Mean from the following Data.

85, 70, 15, 75, 500, 8, 45, 250, 40, 36.

Sol :-

x	$\log x$
85	1.9294
70	1.8450
15	1.1760
75	1.8750
500	2.6989
8	0.9030
45	1.6532
250	2.3979

Here,

N = 10.

40	1.6020
36	1.5563
	$\Sigma \log x = 17.6367$

$$G.M. = A.L\left[\frac{\Sigma \log x}{N}\right]$$

$$= A.L\left[\frac{17.6367}{10}\right]$$

$$= A.L[1.76367]$$

$$\boxed{\therefore G.M. = 58.04}$$

Discrete Series :-

1) Calculate Geometric mean for the following data.

x :	8	9	10	11	12	13	14
f :	11	8	6	9	7	3	1

Sol:-

x	f	logx	flogx
8	11	0.9030	9.933
9	8	0.9542	7.6336
10	6	1	6
11	9	1.0413	9.3717
12	7	1.0791	7.5537
13	3	1.1139	3.3417
14	1	1.1461	1.1461
	$\Sigma f = 45$		$\Sigma f\log x = 44.9798$

$$G.M. = A.L\left[\frac{\Sigma f\log x}{\Sigma f}\right] \Rightarrow A.L\left[\frac{44.9798}{45}\right] \Rightarrow A.L[0.9995]$$

Subject : Date :

Title of the test case :

Case study No. : Page No. :

$\boxed{\therefore G.M = 9.9884}$

Continuous Series :-

1) Calculate Geometric Mean for the following distribution.

class Intervals (x) : 0-10 10-20 20-30. 30-40 40-50

Frequency (f) : 5 7 15 25 8

Sol :-

class Interval(x)	Frequency (f)	Midvalues (m)	log m	flog m
0 - 10	5	$\frac{0+10}{2} = 5$	0.6989	3.4945
10 - 20	7	15	1.1760	8.232
20 - 30	15	25	1.3979	20.9685
30 - 40	25	35	1.5440	38.6
40 - 50	8	45	1.6532	13.2256
	$\Sigma f = 60$			$\Sigma f\log m = 84.5206$

$$\therefore GM = A.L\left[\frac{\Sigma f\log m}{\Sigma f}\right]$$

$$= A.L\left[\frac{84.5206}{60}\right]$$

$$= A.L\left[1.4086\right]$$

$\boxed{\therefore G.M = 25.6212}$

Harmonic Mean (H.M) :-

The Harmonic mean is based on the reciprocals of numbers averaged. It is defined as the reciprocal of the arithmetic mean of the individual observations.

In Individual Series :- $H.M = \frac{N}{\Sigma(1/x)}$

In Discrete series :- $H.M = \frac{N}{\Sigma(f/x)}$

In continuous Series :- $H.M = \frac{N}{\Sigma(f/m)}$

Individual Series :-

1) Find Harmonic mean for the following distribution.

2574, 475, 75, 5, 0.8, 0.08, 0.005, 0.0009.

Sol:-

x	$1/x$
2574	0.0003
475	0.0021
75	0.0133
5	0.2
0.8	1.25
0.08	12.5
0.005	200
0.0009	1111.111
	$\Sigma(1/x) = 1325.07$

$N = 8.$

$$H.M. = \frac{N}{\Sigma(1/x)} = \frac{8}{1325.07}$$

$$\boxed{\therefore H.M = 0.006}$$

Subject :
Date :
Title of the test case :
Case study No. :
Page No. :

Discrete Series :-

1) Calculate Harmonic mean for the following distribution

Marks :	10	20	25	40	50
No. of students :	20	30	50	15	5

Marks (x)	No. of Students (f)	f/x.
10	20	20/10 = 2
20	30	30/20 = 1.5
25	50	50/25 = 2.
40	15	15/40 = 0.375
50	5	5/50 = 0.1
	N = 120.	Σ(f/x) = 5.975

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

$$= \frac{120}{5.975}$$

$$= 20.0836.$$

$$\boxed{\therefore H.M = 20.0836}$$

Continuous Series :-

1) Calculate Harmonic mean from the following distribution.

Marks (x) :	10-20	20-30	30-40	40-50	50-60
Frequency (f) :	4	6	10	7	3

Marks(x)	Midvalues (m)	frequency(f)	f/m
10-20	15	4	4/15 = 0.2666
20-30	25	6	0.24
30-40	35	10	0.2857
40-50	45	7	0.1555
50-60	55	3	0.0545
		N = 30	Σ(f/m) = 1.0023

$$\therefore H.M = \frac{N}{\Sigma(f/m)}$$

$$= \frac{30}{1.0023}$$

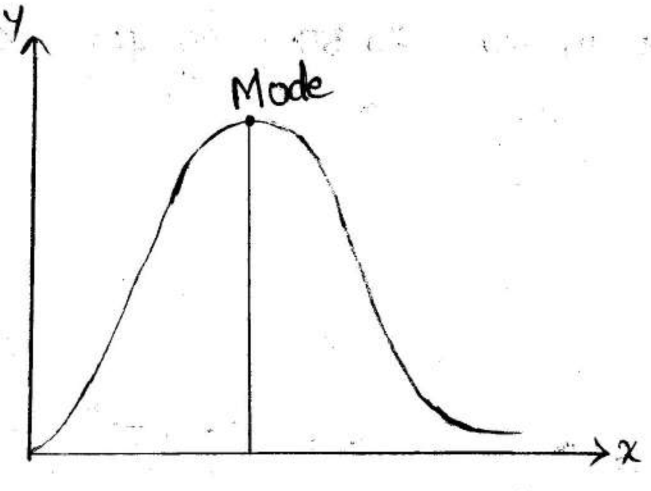
$$= 29.9311...$$

$$\boxed{\therefore H.M = 29.9311}$$

MODE :-

The mode (or) model value is a value in the given series of observations which occurs the greatest frequency.

Graphically :-



The value of the variable at which the curve reaches a maximum is called the "mode".

Individual Series:-

* Arrange in Ascending Order.

* Note the terms that occuring maximum number of values then the term is "mode."

Problem:-

1) Find the mode from the following data.

12, 14, 16, 18, 26, 16, 20, 16, 11, 12, 16, 15, 20, 24.

Sol:- The Ascending Order of the given data is

$11, \boxed{12, 12}, 14, 15, \boxed{16, 16, 16, 16}, 18, \boxed{20, 20}, 24, 26.$

Here, repeating terms are 12, 16, 20

12 - 2 times

16 - 4 times.

20 - 2 times.

Maximum number of repeating term is 16. i.e., 4 times.

$$\boxed{\therefore \text{Mode} = 16.}$$

Discrete Series:-

In discrete series, mode is known by inspection method i.e., the variable which is having highest frequency is called "Mode."

1) Find mode for the following distribution.

x :	4	7	11	16	25
f :	3	9	14	21	13

Sol:-

x :	4	7	11	16	25
f :	3	9	14	21	13

Highest frequency = 21

Highest frequency corresponding variable = 16.

∴ Mode = 16.

The highest frequency having the variable is 16.

$\boxed{\therefore \text{Mode} = 16.}$

In Continuous Series :-

Here we are using the formula.

$$\boxed{\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times C.I.}$$

Where, L = Lower limit of the class Interval.

f_1 = frequency to the class of mode/model value.

f_0 = frequency before preceding value of the model interval.

f_2 = frequency after succeding value of the model interval.

C.I = Length of the class Interval.

Subject	:	Date	:
Title of the test case	:		
Case study No.	:	Page No.	:

1) Calculate mode from the following series.

class Interval :	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency :	4	13	21	44	33	22	7

Sol:-

class Interval	Frequency
0-10	4
10-20	13
20-30	21 - f_0
L $\boxed{30}$-40	44 - f_1
40-50	33 - f_2
50-60	22
60-70	7
	N = 144

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times C.I$$

$$= 30 + \frac{44-21}{2(44)+21-33} \times 10$$

$$= 30 + \frac{23}{88-54} \times 10$$

$$= 30 + \frac{23}{34} \times 10 \Rightarrow 30 + 0.6764 \times 10$$

$$\Rightarrow 30 + 6.764$$

$$\Rightarrow 36.764.$$

$$\boxed{\therefore \text{Mode} = 36.764}$$

MEDIAN :-

Median is a value that divides the series into 2 equal parts. In some cases median is the no. of terms is less than the median (or) the no. of terms is more than the median (or) equal the median. Median is standard by the following series.

Individual Series :-

* Arrange the given data in Ascending Order.

* In individual series 2 cases are existed based on the observations even (or) odd.

Case 1 :-

The number of terms in given data are odd number then we have to choose middle term is the "Median".

1) The Income of five employees are given find median for the given data.

5900, 6950, 7020, 7200, 7280.

Sol :- 5900, 6950, 7020, 7200, 7280.

Median = Middle term.

∴ Median = 7020.

Case 2 :-

The number of terms in given data are even number then we have to choose the value of median is sum of the middle two terms divided by 2.

1) Find Median for the following Series.

20, 68, 50, 15, 35, 98, 33, 44, 56, 64.

Sol:- Arrange in Ascending Order.

15, 20, 33, 35, 44, 50, 56, 64, 68, 96.

$$\text{Median} = \frac{N+1}{2} = \frac{10+1}{2} = \frac{11}{2} = 5.5$$

= 5th and 6th terms are the middle values.

$$= \frac{44+50}{2} = \frac{94}{2} = 47.$$

$$\boxed{\therefore \text{Median} = 47.}$$

Discrete Series :-

* Arrange the data in Ascending order.
* Find cummulative frequencies of the given frequencies.
* Apply the formula median = $\frac{N+1}{2}$th item.
* Now look at the cummulative frequency column and find the total which is equal to $\frac{N+1}{2}$ (or) next highest determined value of the variable to the corresponding cummulative frequency. That gives the value of median.

1) From the following data find the value of median.

Income :	4000	4500	5800	5060	6600	5380
No. of Persons :	24	26	16	20	6	30

Sol: Arrange the given data in Ascending Order.

Income (x)	No. of Persons (f)	cummulative Frequency.
4000	24	24
4500	26	50 = 24+26.
5060	20	70 = 50+20 — Median
5380	30	100 = 70+30
5800	16	116 = 100+16
6600	6	122 = 116+6
	N = 122.	

$$\text{Median} = \frac{N+1}{2}^{\text{th item}} = \frac{122+1}{2} = \frac{123}{2} = 61.5$$

$$\boxed{\therefore \text{Median} = 5060.}$$

Continuous Series :-

* Arrange data in Ascending order.

* Calculate the cummulative frequencies.

* Apply the formula median $= L + \dfrac{\frac{N}{2} - C.f}{f} \times C.I.$

where, L = Lower limit of the median class Interval.

C.f = cummulative frequency of the preceding the value of the median class.

Subject :
Title of the test case :
Case study No. :
Date :
Page No. :

f= frequency of the median class.

C.I = Length of the class Interval.

1) calculate median for the following distribution.

Marks :	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
No. of students :	7	15	24	31	42	30	26	15	10

Soln

Given,

Marks (x)	No. of students (f)	C.f.
5-10	7	7
10-15	15	22
15-20	24	46
20-25	31	77
25-30	42	119
30-35	30	149
35-40	26	175
40-45	15	190
45-50	10	200.
	N = 200	

calculate

$$\frac{N}{2} = \frac{200}{2} = 100$$

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times C.I$$

$$= 25 + \frac{100-77}{42} \times 5 \Rightarrow 25 + \frac{23}{42} \times 5 \Rightarrow 25 + 2.7380$$

Prepared By,
M. Navaneeth Kumar Reddy
Asst. Prof

$\Rightarrow 27.7380 \Rightarrow 27.73$.

$\boxed{\therefore \text{Median} = 27.73.}$

Measure Of Dispersion :-

According to "A.L. Bowley", Dispersion is the measure of the variation of the items.

Properties of a good measure of Dispersion :-

* It should be simple to understand.
* It should be easy to calculate.
* It should be strongly defined.
* It should be based on each & every item of distribution.
* It should be used ~~for further~~ algebric expressions.

Methods of studying Variation :-

* Range.
* Inter quartile deviation/quartile deviation.
* Mean deviation.
* standard deviation.
* co-efficient of variation.

Range :-

Range is the simplest method of studying dispersion. It is defined as the difference between the value of the smallest item and the value of the largest item included in the distribution.

Symbolically, $\boxed{\text{Range} = L - S.}$

where, L= largest value
S= smallest value.

The relative measure corresponding to range called "co-efficient of range" obtained by applying the formula.

$$\boxed{\text{co-efficient of range} = \frac{L-S}{L+S}.}$$

Measure of Dispersion:-

According to A.L. Bowlany "Dispersion is the measure of the variation of the items"

According to Brooks & Dick, "Dispersion is the degree of the scatter (or) variation of the variable about a central value."

Significance of Measuring Variation:-

* To determine the repliability of an average.
* To serve as a basis for the control of the variability.
* To compare two (or) more series with regard to their variability.
* To facilitate the use of other statistical measures.

Properties of a Good Measure of Variation:-

* It should be simple to understand.
* It should be easy to compute.
* It should be strongly define.
* It should be based on each & every item of the

distribution.

- * It should be amenable to further algebraic treatment.
- * It should have sampling stability.
- * It should be unduly affected by the extreme items.

Methods of Studying Variation :-

The following are the important methods of studying variation.

- * The Range.
- * The Interquartile Range & the quartile deviation.
- * The mean deviation | Average deviation.
- * The standard deviation.
- * co-efficient of variation.

Individual Series :-

1) Find the range & co-efficient of range for the following observations.

10, 8, 5, 10, 9, 14, 7.

A) Largest value, $L = 14$.

Smallest value, $S = 5$

$$\text{range} = L - S$$
$$= 14 - 5$$
$$= 9.$$

$\therefore$ co-efficient of range $= \dfrac{L-S}{L+S}$

$$= \frac{14-5}{14+5}$$
$$= \frac{9}{19}$$
$$= 0.473$$

Subject :
Date :
Title of the test case :
Case study No. :
Page No. :

Discrete Series :-

1) Find the range and co-efficient of range for the following data.

x :	11	18	29	33	37	39	40	42	43
f :	100	105	91	82	61	32	70	88	67.

Sol:- Largest value, L = 43

Smallest value, s = 11

Range = L-S = 43-11 = 32.

∴ co-efficient of range $= \frac{L-S}{L+S} = \frac{43-11}{43+11} = \frac{32}{54} = 0.5925.$

Continuous Series :-

1) Find range & co-efficient of range for the following data.

Marks :	10-20	20-30	30-40	40-50	50-60
No. of students :	22	28	44	43	49.

Sol:- Largest value, L = 60

Smallest value, s = 10

Range = L-S = 60-10 = 50.

co-efficient of range $= \frac{L-S}{L+S} = \frac{60-10}{60+10} = \frac{50}{70} = 0.7142.$

Inter quartile deviation / Quartile deviation :-

The inter quartile deviation represent the difference between the third quartile to the first quartile Symbolically, Interquartile range = $Q_3 - Q_1$, and.

quartile deviation is quartile range divided by '2.'

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

$$Q_1 = \text{Size of } \left(\frac{n+1}{4}\right) \text{ item}$$

$$Q_3 = \text{Size of } 3\left(\frac{n+1}{4}\right) \text{ item.}$$

Individual Series :-

1) Find the quartile deviation and co-efficient of Q.D for the following data.

8, 10, 14, 22, 26, 28, 30, 36, 44, 59, 64.

Sol:- Arrange the data in Ascending order.

8, 10, 14, 22, 26, 28, 30, 36, 44, 59, 64

N = 11. (or) n = 11.

$$Q_1 = \text{Size of } \left(\frac{n+1}{4}\right)$$

$$= \text{Size of } \left(\frac{11+1}{4}\right)$$

$$= \text{Size of } \left(\frac{12}{4}\right)$$

$$= \text{Size of (3) item.}$$

$$Q_1 = 14.$$

$$Q_3 = \text{Size of } 3\left(\frac{n+1}{4}\right)$$

$$= \text{Size of } 3\left(\frac{11+1}{4}\right)$$

$$= \text{Size of} 3\left(\frac{12}{4}\right)$$

$$= \text{Size of } 3(3)$$

$$= \text{Size of } 9^{th} \text{ item.}$$

$$Q_3 = 44.$$

$$Q.D = \frac{Q_3-Q_1}{2} = \frac{44-14}{2} = \frac{30}{2} = 15.$$

$$\text{coefficient of } Q.D = \frac{Q_3-Q_1}{Q_3+Q_1} = \frac{44-14}{44+14} = \frac{30}{58} = 0.517$$

Subject : Date :

Title of the test case :

Case study No. : Page No. :

Discrete Series :-

1) Find Q.D & co-efficient of Q.D for the following data.

x : 40 45 50 55 60 65 70 75 80

f : 20. 36 44 50 80 30 30 16 14.

Sol:-

x	f	c.f
40	20	20
45	36.	56
$\boxed{50}$	44	100 ←
55	50	150
60	80	230
$\boxed{65}$	30	260 ←
70	30	290
75	16	306
80	14	320
	N = 320	

$Q_1 = \text{Size of } \left(\frac{n+1}{4}\right)$

$= \left(\frac{320+1}{4}\right)$

$= \frac{321}{4}$

$Q_1 = \text{Size of } (80.25)$

$Q_1 = 50.$

$Q_3 = \text{Size of } 3\left(\frac{n+1}{4}\right)$

$= 3(80.25)$

$= 240.75$

$Q_3 = \text{Size of } (240.75)$

$Q_3 = 65$

$Q.D = \frac{Q_3 - Q_1}{2} = \frac{65-50}{2} = \frac{15}{2} = 7.5$

$\therefore$ co-efficient of $Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{65-50}{65+50} = \frac{15}{115} = 0.1309$

Continuous Series :-

In continuous series, $Q_1 = L_1 + \dfrac{\frac{N}{4} - c.f_1}{f_1} \times C.I.$

$$Q_3 = L_3 + \frac{3(\frac{N}{4}) - cf_3}{f_3} \times C.I.$$

where, L_1, L_3 = lower limit of the class intervals.

N = Sum of the frequency.

c.f = cummulative frequencies preceeding value of C.I.

f_1, f_3 = frequency of C.I.

C.I = Length of class Interval.

1) Find Q.D & co-efficient of Q.D for the following data.

x:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80.
f:	5	8	7	12	28	20	10	10

Sol:-

x	f	c.f
0-10	5	5
10-20	8	13
20-30	7	20 cf_1
L_1 $\boxed{30}$-40	12 f_1	32 cf_2
40-50	28 f_2	60 cf_3
L_3 $\boxed{50}$-60	20 f_3	80
60-70	10	90
70-80	10	100
	N =100	

$\frac{N}{4} = \frac{100}{4} = 25.$

$\frac{N}{2} = \frac{100}{2} = 50.$

$Q_1 = L_1 + \dfrac{\frac{N}{4} - cf_1}{f_1} \times C.I.$

$= 30 + \dfrac{25-20}{12} \times 10$

$= 30 + \frac{5}{12} \times 10$

$= 30 + 0.4166 \times 10$

$Q_3 = L_3 + \dfrac{3(\frac{N}{4}) - cf_3}{f_3} \times C.I.$

$= 50 + \dfrac{75-60}{20} \times 10$

$= 50 + \frac{15}{20} \times 10$

$= 50 + 7.5$

Subject	:	Date	:
Title of the test case	:		
Case study No.	:	Page No.	:

$= 30 + 4.166$

$Q_1 = 34.166$

$Q_3 = 57.5$

$$Q.D = \frac{Q_3 - Q_1}{2} = \frac{57.5 - 34.166}{2} = \frac{23.39}{2} = 11.67$$

$$\text{co-efficient of } Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{57.5 - 34.166}{57.5 + 34.166} = \frac{23.34}{91.66}$$

$$= 0.2546$$

Mean Deviation:-

Mean deviation is also known as the average deviation. It is the difference between the items in a distribution and the median mean (or) mode of that series.

In Individual Series:-

Mean deviation, $M.D = \frac{\Sigma D}{N}$ here $D = |x - \bar{x}|$

mean, $D = |x - \text{mean}|$

median, $D = |x - \text{median}|$

mode, $D = |x - \text{mode}|$

$$\therefore \text{co-efficient of mean deviation} = \frac{\text{Mean deviation}}{\text{Mode / Median / Mean.}}$$

1) Calculate the mean deviation and co-efficient of mean deviation with the help of mean, mode, median from the following data.

4, 7, 7, 7, 9, 9, 10, 12, 15.

Sol:- **With Mean :-**

Mean

x	$D = \lvert x - \bar{x} \rvert$
4	$\lvert 4-9 \rvert = 5$
7	$\lvert 7-9 \rvert = 2$
7	$\lvert 7-9 \rvert = 2$
7	$\lvert 7-9 \rvert = 2$
9	$\lvert 9-9 \rvert = 0$
9	$\lvert 9-9 \rvert = 0$
10	$\lvert 10-9 \rvert = 1$
12	$\lvert 12-9 \rvert = 3$
15	$\lvert 15-9 \rvert = 6$
$\Sigma x = 80$	$\Sigma D = 21.$

$$\text{Mean, } \bar{x} = \frac{\Sigma x}{N} = \frac{80}{9} = 8.88$$

$$\bar{x} \approx 9.$$

$$\therefore \text{Mean deviation, M.D} = \frac{\Sigma D}{N} = \frac{21}{9} = 2.33.$$

$$\therefore \text{co-efficient of mean deviation} = \frac{\text{M.D}}{\text{mean}} = \frac{2.33}{9} = 0.259$$

With Mode :- $N=9$.

Mode

x	$D = \lvert x - \bar{x} \rvert$
4	$\lvert 4-7 \rvert = 3$
7	$\lvert 7-7 \rvert = 0$
7	$\lvert 7-7 \rvert = 0.$
7	$\lvert 7-7 \rvert = 0$
9	$\lvert 9-7 \rvert = 2$
9	$\lvert 9-7 \rvert = 2$
10	$\lvert 10-7 \rvert = 3$
12	$\lvert 12-7 \rvert = 5$
15	$\lvert 15-7 \rvert = 8$
	$\Sigma D = 23.$

$\therefore$ Mode, $\bar{x} = 7$. (highest number of repeating term.)

Subject : Date :

Title of the test case :

Case study No. : Page No. :

∴ Mean deviation, $M.D = \frac{\Sigma D}{N}$

$= \frac{23}{9}$

$= 2.55.$

∴ co-efficient of mean deviation $= \frac{\text{Mean deviation}}{\text{Mode}}$

$= \frac{2.55}{7}$

$= 0.364.$

With Median :-

Median

x	$D = \|x-\bar{x}\|$
4	$\|4-9\| = 5$
7	$\|7-9\| = 2$
7	$\|7-9\| = 2$
7	$\|7-9\| = 2$
9	$\|9-9\| = 0$
9	$\|9-9\| = 0$
10	$\|10-9\| = 1$
12	$\|12-9\| = 3$
15	$\|15-9\| = 6$
	$\Sigma D = 21.$

∴ Median = 9, (here no. of terms is 9. So, odd terms existed. So, middle term is called "Median".)

∴ Mean deviation, $M.D = \frac{\Sigma D}{N} = \frac{21}{9} = 2.33.$

∴ co-efficient of mean deviation $= \frac{M.D}{\text{Median}}$

$= \frac{2.33}{9}$

$= 0.259$

Discrete Series:-

$$\text{Mean deviation} = \frac{\Sigma FD}{N}$$

where, f = frequency

$D = |x-\bar{x}|$ → mean, median, mode.

N = Sum of the frequency.

$\therefore$ co-efficient of mean deviation = $\dfrac{\text{Mean deviation}}{\text{Mean}|\text{Mode}|\text{Median}}$.

1) Calculate Mean deviation and co-efficient of mean deviation for the following data with the help of mean, mode, median?

x : 10 11 12 13 14

f : 3 12 18 12 3.

Sol:- **With Mean:-**

x	f	fx	$D=\|x-\bar{x}\|$	fD
10	3	30	|10-12| = 2	3x2 = 6
11	12	132	|11-12| = 1	12x1 = 12
12	18	216	|12-12| = 0	18x0 = 0
13	12	156	|13-12| = 1	12x1 = 12
14	3	42	|14-12| = 2	14x2 = 28
	N=48	Σfx = 576		ΣfD = 56

$$\text{Mean}, \bar{x} = \frac{\Sigma fx}{N} = \frac{576}{48}$$

$\bar{x} = 12$

$\therefore$ Mean deviation, $M.D = \dfrac{\Sigma fd}{N} = \dfrac{56}{48} = 1.1666.$

$\therefore$ co-efficient of mean deviation $= \dfrac{M.D}{\text{mean}}$

$= \dfrac{1.1666}{12}$

$= 0.0972$.

Subject : Date :

Title of the test case :

Case study No. : Page No. :

With Mode :-

x	f	$D=\|x-\bar{x}\|$	fD
10	3	$\|10-12\|=2$	6
11	12	$\|11-12\|=1$	12
12 - mode	18	$\|12-12\|=0$	0
13	12	$\|13-12\|=1$	12
14	3	$\|14-12\|=2$	6
	$N=48$		$\Sigma fD=36$

$\therefore$ Mode = highest frequency variable $=12$.

$\therefore$ Mean deviation, $M \cdot D = \frac{\Sigma fd}{N} = \frac{36}{48} = 0.75$.

$\therefore$ co-efficient of $M \cdot D = \frac{\text{Mean deviation}}{\text{Mode}} = \frac{0.75}{12} = 0.06$

With Median :-

x	f	cf	$d=\|x-\bar{x}\|$	fd
10	3	3	$\|10-12\|=2$	6
11	12	15	$\|11-12\|=1$	12
12	18	33	$\|12-12\|=0$	0
13	12	45	$\|13-12\|=1$	12
14	3	48	$\|14-12\|=2$	6
				$\Sigma fd=36$

calculate $\frac{N+1}{2}$

$$=\frac{48+1}{2}$$

$$=\frac{49}{2}$$

$$=24.5$$

$\therefore$ Median $\bar{x}=12$.

Mean deviation, $M \cdot D = \frac{\Sigma fD}{N} = \frac{36}{48} = 0.75$.

$\therefore$ co-efficient of $M \cdot D = \frac{\text{Mean deviation}}{\text{Median}} = \frac{0.75}{12} = 0.06$

Continuous Series:-

Mean deviation $= \frac{\Sigma fd}{N}$

where, d = deviation $|x-\bar{x}|$

$\bar{x}$ = mean/mode/median.

N = sum of the frequency.

co-efficient of M.D $= \frac{M.D}{Mean/Mode/Median}$.

1) calculate mean deviation and co-efficient of mean deviation from the following data with the help of mean?

class (x):	0-10	10-20	20-30	30-40	40-50	50-60.
Frequency(f):	5	8	7	12	28	20

Sol:-

x	f	m	fm	$d=\lvert x-\bar{x}\rvert$	fd
0-10	5	5	25	34	170
10-20	8	15	120	24	192
20-30	7	25	175	14	98
30-40	12	35	420	4	48
40-50	28	45	1260	6	168
50-60	20	55	1100	16	320
	N=80		Σfm=3100		Σfd= 996.

Mean, $\bar{x} = \frac{\Sigma fm}{N} = \frac{3100}{80} = 38.75$

$\boxed{\bar{x} \simeq 39.}$

Mean deviation, $M.D = \frac{\Sigma fd}{N} = \frac{996}{80} = 12.45.$

∴ co-efficient of Mean deviation $= \frac{\text{Mean deviation}}{\text{Mean}}$

$= \frac{12.45}{39} = 0.3192$

Standard Deviation :-

The standard deviation concept was introduced by "Karl Pearson" in the year 1823. It is mostly used to measure the studying of dispession. standard deviation is also known as "Root mean square deviation". For this reason standard deviation is square root of the mean square deviation from the arithmetic mean. Symbolically, it is denoted by "σ".

$$S.D,\ \sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

Note :- The relative measure of standard deviation is called "variance". Symbolically it is denoted by "σ^2".

In Individual Series :-

$$S.D,\ \sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

where, d = deviation calculated from mean $\bar{x}$

$d = x - \bar{x}$.

N = Total no. of variables.

Variance $= (\sigma)^2$.

1) Calculate standard deviation and variance from the following data.

120, 100, 160, 100, 220, 130, 150, 170, 150, 200.

Sol:- $N = 10$.

x	$d = (x - \bar{x})$	d^2
120	120-150 = -30	$(-30)^2 = 900$
100	100-150 = -50	$(-50)^2 = 2500$
160	160-150 = 10	$(10)^2 = 100$
100	100-150 = -50	$(-50)^2 = 2500$
220	220-150 = 70	$(70)^2 = 4900$
130	130-150 = -20	$(-20)^2 = 400$
150	150-150 = 0	$(0)^2 = 0$
170	170-150 = 20	$(20)^2 = 400$
150	150-150 = 0	$(0)^2 = 0$
200	200-150 = 50	$(50)^2 = 2500$
$\Sigma x = 1500$	$\Sigma d = 0.$	$\Sigma d^2 = 14,200$

Mean $\bar{x} = \frac{\Sigma x}{N}$

$= \frac{1500}{10}$

$\bar{x} = 150.$

standard deviation, $\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2} = \sqrt{\frac{14200}{10} - \left(\frac{0}{10}\right)^2}$

$= \sqrt{1420 - 0} = \sqrt{1420} = 37.68$

$\sigma = 37.68.$

Variance $= (\sigma)^2 = (37.68)^2 = 1420.$

$\sigma^2 = 1420$

Discrete Series :-

standard deviation, $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$

fd = frequency with deviation.

N = sum of the frequencies.

1) Calculate S.D and variance from the following Data.

x :	10	20	30	40	50	60
f :	8	12	20	10	7	3

Subject :   Date :

Title of the test case :

Case study No. :   Page No. :

Sol:

				31		
x	f	fx	$d\ (x-\bar{x})$	d^2	fd	fd^2
10	8	80	-21	441	-168	3528
20	12	240	-11	121	-132	1452
30	20	600	-1	1	-20	20
40	10	400	9	81	90	810
50	7	350	19	361	133	2527
60	3	180	29	841	87	2523
	$N=60$	$\Sigma fx=1850$			$\Sigma fd=-10.$	$\Sigma fd^2=10{,}860.$

Mean, $\bar{x} = \frac{\Sigma fx}{N} = \frac{1850}{60} = 30.8$

$\boxed{\bar{x} \approx 31.}$

standard deviation, S.D, $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$

$= \sqrt{\frac{10860}{10} - \left(\frac{-10}{60}\right)^2}$

$= \sqrt{181 - (0.166)^2}$

$= \sqrt{181 - 0.027}$

$= \sqrt{180.973}$

$\sigma = 13.45.$

· Variance, $\sigma^2 = (13.45)^2 = 180.973.$

b) Continuous Series :-

standard deviation, S.D, $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times C.$

where, fd = frequency with mid values deviation.

C = length of the class Interval.

1) Calculate standard deviation and variance from the following data.

class (x):	0-10	10-20	20-30	30-40	40-50	50-60
Frequency(f):	8	12	20	10	7	3

Sol: (i)

x	f	m	fm	$\overset{26}{d(x-\bar{x})}$	d^2	fd	fd^2
0-10	8	5	40	-21	441	-168	3528
10-20	12	15	180	-11	121	-132	1452
20-30	20	25	500	-1	1	-20	20
30-40	10	35	350	9	81	90	$8\overset{1}{\cancel{4}}0$
40-50	7	45	315	19	361	133	2527
50-60	3	55	165	29	841	87	2523
	N=60		$\Sigma fm=1550$			$\Sigma fd=-10$	$\Sigma fd^2=10860.$

$$\text{Mean}, \bar{x} = \frac{\Sigma fm}{N} = \frac{1550}{60} = 25.83$$

$$\boxed{\therefore \bar{x} \approx 26.}$$

- Standard deviation, S.D, $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times C$

$$= \sqrt{\frac{10860}{60} - \left(\frac{-10}{60}\right)^2} \times 10$$

$$= \sqrt{181 - (0.166)^2} \times 10 = \sqrt{181 - 0.027} \times 10$$

$$= \sqrt{180.973} \times 10$$

$$= 13.45 \times 10$$

$$\sigma = 134.5.$$

$$\text{Variance} = \sigma^2 = (134.5)^2$$

$$\sigma^2 = 18090.25.$$

Subject :
Title of the test case :
Case study No. :
Date :
Page No. :

Co-efficient of Variation :-

Standard deviation is discussed absolute measure of dispersion the corresponding relative measure is called "co-efficient of Variation". This method is developed by "Karl Pearson." This is mostly used to calculate the relative measure is called "Variation". It is used in such probleme where we want to compare the variability of two (or) more series.

$$\therefore \text{co-efficient of variation} = \frac{\sigma}{\bar{x}} \times 100.$$

where, σ = standard deviation.

$\bar{x}$ = mean.

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

1) From the prices of shares of 'x' and 'y' given below. Find out which is more stable in value.

x : 35 54 52 53 56 58 52 50 51 49

y : 108 107 105 105 106 107 104 103 104 101.

Solution

x	$X (x-\bar{x})$	X^2	y	$Y (y-\bar{y})$	Y^2
35	-16	256	108	3	9
54	3	9	107	2	4
52	1	1	105	0	0
53	2	4	105	0	0
56	5	25	106	1	1
58	7	49	107	2	4
52	1	1	104	-1	1
50	-1	1	103	-2	4
51	0	0	104	1	1
49	-2	4	101	-4	16
$\Sigma x = 510$		$\Sigma X^2 = 350$	$\Sigma y = 1050$		$\Sigma Y^2 = 40.$

Mean, $\bar{x} = \frac{\Sigma x}{N} = \frac{510}{10} = 51.$

$\boxed{\bar{x} = 51.}$

Mean, $\bar{y} = \frac{\Sigma y}{N} = \frac{1050}{10} = 105$

$\boxed{\bar{y} = 105.}$

co-efficient of variation in $x = \frac{\sigma}{\bar{x}} \times 100$

$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{350}{10}} = \sqrt{35} = 5.91.$

co-efficient of variation, $x = \frac{\sigma}{\bar{x}} \times 100$

$= \frac{5.91}{51} \times 100 = 0.1158 \times 100$

$= 11.58.$

co-efficient of variation in $y = \frac{\sigma}{\bar{y}} \times 100$

$\sigma = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{40}{10}} = \sqrt{4} = 2.$

co-efficient of variation, $y = \frac{\sigma}{\bar{y}} \times 100 = \frac{2}{105} \times 100 = 0.0190 \times 100$

$= 1.90.$

∴ Here, co-efficient of variation in 'x' is more when compare to co-efficient of variation in 'y'. So, shares 'y' is more stable to shares 'x'.

Applications of Measure of Central Tendency & Dispersion:-

Central Tendency and dispersion can be used for

* In finance measures of central tendency and dispersion is used as an indicator of the risk involved in an investment. Since, it measures the variability of returns around the expected return from an investment.

Subject	:	Date	:
Title of the test case	:		
Case study No.	:	Page No.	:

* Financial managers can also use expected value and dispersion action to make important inferences from the past data.
* Measures of central Tendency & dispersion is used in determining the variability in sales & earnings.
* The expected returns & its associated standard deviations are used as measure of risk on investment analysis. These measures used in comparing investments such as stocks, bonds & even mutual funds.
* The measures of central tendency and dispersion can be used to analyse sample market survey data, rates of return on a stock and economic data.

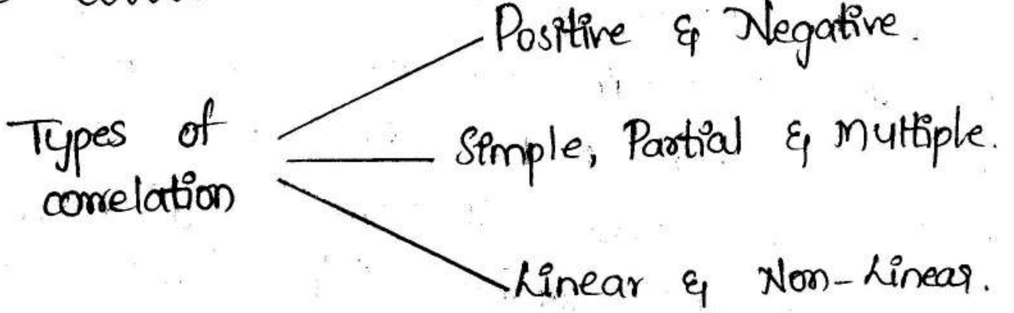
UNIT-2

CORRELATION

Types of Correlation :-

Definition :- A statistical tool used to measure the relationship between two (or) more variables such that the movement in one variable is accompained by the movement of another is called as "Correlation."

Types of Correlation :-



Types of correlation

Positive & Negative.

Simple, Partial & Multiple.

Linear & Non-Linear.

Positive & Negative Correlation :-

Whether the correlation between the variables is positive (or) negative depends on it's direction of change. The correlation is positive when both the variables move in the same direction, i.e., when one variable increases the other on an average also increases and if one variable, decreases the other also decreases. The correlation is said to be negative when both the variables move in the opposite direction, i.e., when one variable increases the other decreases & vice versa.

Simple, Partial And Multiple Correlations :-

Whether the correlation is Simple, partial (or) multiple depends on the number of variables studied.

The correlation is said to be simple when only two variables are studied. The correlation is either multiple or partial when three (or) more variables are studied. the correlation is said to be Multiple when three variables are studied simultaneously. such as, if we want to study the relationship between the yield of wheat per acre and the amount of fertilizers and rainfall used, then it is a problem of multiple correlations.

Whereas, in the case of a partial correlation we study more than two variables, but consider only two among them that would be influencing each other such that the effect of the other influencing variable is kept constant. Such as, in the above example, if we study the relationship between the yield and fertilizers used during the periods when certain average temperature existed, then it is a problem of partial correlation.

Linear & Non-Linear (curvilinear) Correlation:-

Whether the correlation between the variables is linear (or) non-linear depends on the constancy of ratio of change between the variables. The correlation is said to be linear when the amount of change in one variable to the amount of change in another variable tends to bear a constant ratio. For example, from the values of two variables given below, it is clear that the ratio of change between the variables is the Same :

X : 10 20 30 40 50.

Y : 20 40 60 80 100.

The correlation is called as "Non-Linear (or) curvilinear" when the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable. For example, if the amount of fertilizers is doubled the yield of wheat would not be necessarily be doubled.

Thus, there are three most important types of correlation classified on the basis of movement, number and the ratio of change between the variables. The researcher must study these carefully to determine the correlation methods to be used to identify the extent to which the variables are correlated.

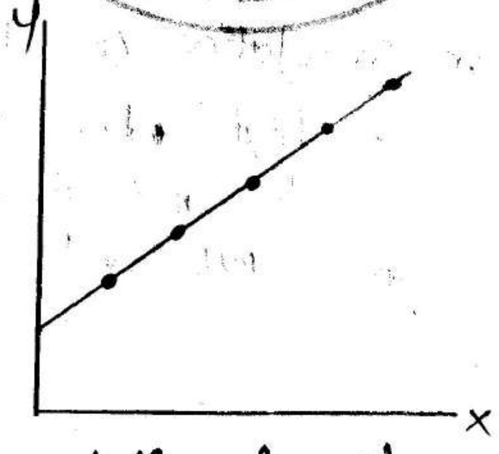
Methods of Determining Correlation :-

Definition :- The scatter diagram method is the simplest method to study the correlation between two variables wherein the values for each pair of a variable is plotted on a graph in the form of dots thereby obtaining as many points as the number of observations. Then by looking at the scatter of several points, the degree of correlation is ascertained.

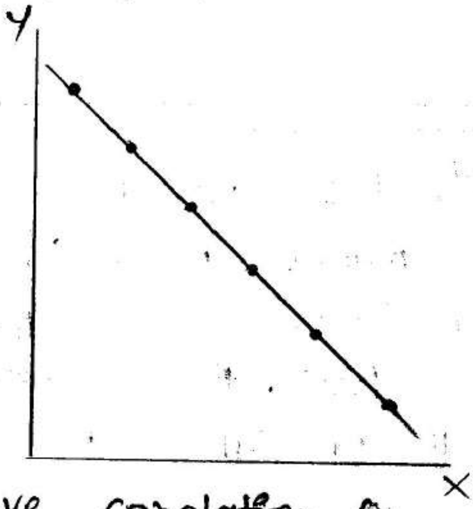
The degree to which the variables are related to each other depends on the manner in which the points are scattered over the chart. The more the points plotted are scattered over the chart, the lesser is the degree of correlation between the variables. The more the points plotted are closer to the line, the higher is the degree of correlation. The degree of correlation is denoted by "r".

The following types of scatter diagrams tell about the degree of correlation between variable x and variable y.

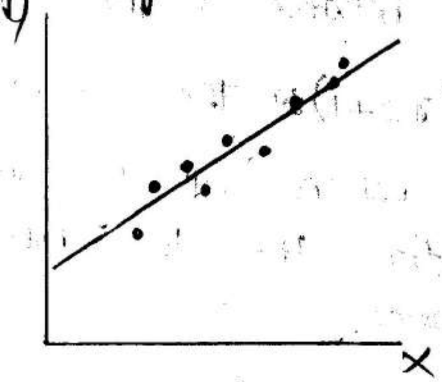
Positive Correlation (r=+1):- The correlation is said to be perfectly positive when all the points lie on the straight line rising from the lower left-hand corner to the upper right-hand corner.



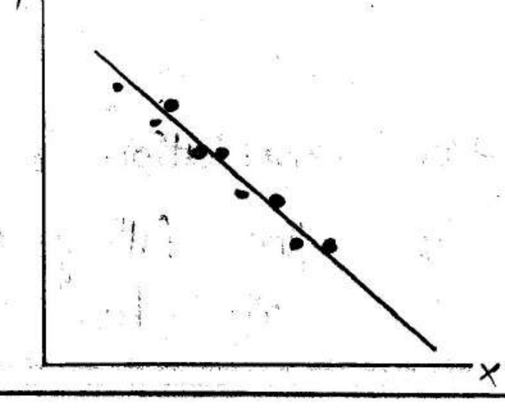
Perfect Negative Correlation (r=-1):- When all the points lie on a straight line falling from the upper left-hand corner to the lower right-hand corner, the variables are said to be negatively correlated.



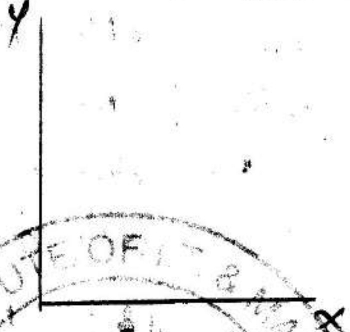
High Degree of +ve correlation (r = + high) :- The degree of correlation is high when the points plotted fall under the narrow band and is said to be positive when these show the rising tendency from the lower left-hand corner to the upper right-hand corner.



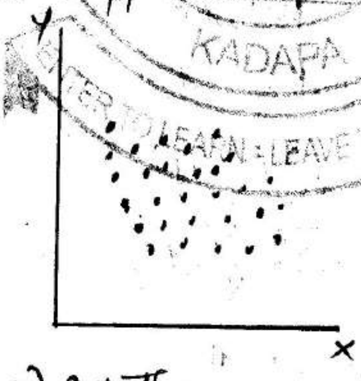
High degree of -ve correlation (r = -high) :- The degree of negative correlation is high when the point plotted fall in the narrow band and show the declining tendency from the upper left-hand corner to the lower right-hand corner.



Low Degree of +ve Corelation (r = +Low) :→ The correlation between the variables is said to be low but positive when the points are highly scattered over the graph & show a rising tendency from the lower left-hand corner to the upper right-hand corner.



Low Degree of -ve Corelation (r = +Low) :→ The degree of correlation is low and negative when the points are scattered over the graph and the show the falling tendency from the upper left-hand corner to the lower right-hand corner.



No Corelation (r=0) :→ The variable is said to be unrelated when the point are haphazardly scattered over the graph and do not show any specific pattern. Here the correlation is absent and hence r=0.

Thus, the scatter diagram method is the simplest device to study the degree of relationship between the variables by plotting the dots to each pair of variable values given. The chart on which the dots are plotted is also called as a "Dotogram."

Karl Pearson's Co-efficient of Corelation :-

Definition :- Karl pearson's co-efficient of correlation is widely used mathematical method where in the numerical expression is used to calculate the degree and direction of the relationship between linear related variables.

Pearson's method, popularly known as a "Pearson Co-efficient of Corelation", is the most extensively used quantitative methods in practice. The co-efficient of correlation is denoted by "r".

If the relationship between two variables x and y is to be ascertained, then the following formula is used:

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

where, $\bar{x}$ = mean of x variable

$\bar{y}$ = mean of y variable.

Properties of Co-efficient of Correlation :-

* The value of the co-efficient of correlation (r) always lies between ±1. such as:

$r=+1$, perfect positive correlation.

$r=-1$, perfect negative correlation.

r=0, no correlation.

* The co-efficient of correlation is independent of the origin & scale. By origin, it means subtracting any non-zero constant from the given value of x and y the value of "r" remains unchanged. By scale it means, there is no effect on the value of "r" if the value of x and y is divided (or) multiplied by any constant.

* The co-efficient of correlation is a geometric mean of two regression co-efficient. Symbolically, it is represented as

$$r = \sqrt{bxy + byx}$$

* co-efficient of correlation is "zero" when the variables x and y are independent. But, however, the converse is not true.

Assumptions of Kaal Pearson's Co-efficient of Correlation:-

* The relationship between the variables is "linear", which means when the two variables are plotted, a straight line is formed by the points plotted.

* There are a large no. of independent causes that affect the variables under study so as to form a Normal Distribution. such as, variables like price, demand, supply, etc. are affected by such factors that the normal distribution is formed.

* The variables are independent of each other.

Note:- The co-efficient of correlation measures not only the magnitude of correlation but also tells the direction, such as, $r = -0.67$, which shows correlation is negative because the sign is "–" and the magnitude is 0.67.

Spearman's Rank Correlation co-efficient:-

Definition:- The Spearman's Rank Correlation Co-efficient is the non-parametric statistical measure used to study the strength of association between the two ranked variables. This method is applied to the ordinal set of numbers, which can be arranged in order, i.e., one after the other so that ranks can be given to each.

In the rank correlation co-efficient method, the ranks are given to each individual on the basis of it's quality (or) quantity, such as ranking starts from position 1st and goes till Nth position for the one ranked last in the group.

The formula to calculate the rank correlation co-efficient is:

$$R = \frac{(1-6\Sigma D^2)}{N(N^2-1)} = \frac{(1-6\Sigma D^2)}{N^3-N}$$

where, R = Rank co-efficient of correlation.

D = Difference of ranks.

N = Number of observations.

The value of R lies between ±1 such as:

R = +1, there is a complete agreement in the order of ranks and move in the same direction.

R = -1, there is a complete agreement in the order of ranks, but are in opposite directions.

R=0, there is no association in the ranks.

While solving for the rank correlation co-efficient one may come across the following problems:

→ Where actual Ranks are given.

→ Where ranks are not given.

→ Equal ranks (or) Tie in Ranks.

Where actual ranks are given:- An individual must follow the following steps to calculate the correlation coefficient:

* First, the difference between the ranks (R1-R2) must be calculated, denoted by D.

* Then, square these differences to remove the negative sign & obtain its sum ΣD^2.

* Apply the formula as shown above.

Where ranks are not given:- In case the ranks are not given, then the individual may assign the rank by taking either the highest value (or) the lowest value as 1. Whatever criteria is being decided the same method should be applied to all the variables.

Equal Ranks (or) Tie in Ranks:- In case the same ranks are assigned to two (or) more entities, then the ranks are assigned on an average basis. such as if two individuals are ranked equal at third position, then the ranks shall be calculated as :

$(3+4)/2 = 3.5$

The formula to calculate the rank correlation co-efficient when there is a tie in the ranks is :-

$$R = 1 - \frac{6[6\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + ----)}{N^3 - N}.$$

where m = number of items whose ranks are common.

Note:- The spearman's rank correlation co-efficient method is applied only when the initial data are in the form of ranks, and N (Number of observations) is fairly small, i.e., not greater than 25 (or) 30.

Key differences between Correlation and Regression:-

The points given below, explains the difference between correlation & regression in detail:

* A statistical measure which determines the co-relationship (or) association of two quantities is known as "Correlation". Regression describes how an independent variable is numerically related to the dependent variable.

* Correlation is used to represent the linear relationship between two variables. On the contrary, regression is used to fit the best line and estimate one variable on the basis of another variable.

Subject :　　　　　　　　　　　　　　Date :

Title of the test case :

Case study No. :　　　　　　　　　　　Page No. :

* In correlation, there is no difference between dependent & independent variables i.e., correlation between x and y is similar to y and x. converselly, the regression of y on x is different from x on y.

* Correlation indicates the strength of association between variables. As opposed to, regression reflects the impact of the unit change in the independent variable on the dependent variable.

* Correlation aims at finding a numerical value that expresses the relationship between variables. Unlike regression whose goal is to predict values of the random variable on the basis of the values of fixed variable.

Meaning of Regression coefficient :-

Regression co-efficient is a statistical measure of the average functional relationship between two (or) more variables. In regression analysis, one variable is considered as dependent and others as independent. Thus, it measures the degree of dependence of one variable on the others. Regression co-efficient was first used for estimating the relationship between the heights of fathers and their sons.

Properties of Regression Co-efficient :-

The important properties of regression co-efficient are given below :

* It is denoted by b.

* It is expressed in terms of original unit of data.

* Between two variables (say x and y), two values of regression co-efficient can be obtained. One will be obtained when we consider x as independent and y as dependent and the other when we consider y as independent and x as dependent. The regression co-efficient of y on x is represented as byx and that of x on y as bxy.

* Both regression co-efficients must have the same sign. If byx is positive, bxy will also be positive & viceversa.

* If one regression co-efficient is greater than unity, then the other regression co-efficient must be lesser than unity.

* The geometric mean between two regression co-efficients is equal to the co-efficient of correlation, r = doubt

* Arithmetic mean of both regression co-efficients is equal to (or) greater than co-efficient of correlation.

(byx + bxy)/2 = equal (or) greater than r.

Regression co-efficients are classified as:

1) Simple, partial and multiple.

2) Positive and negative and

3) Linear and non-linear.

Subject : Date :
Title of the test case :
Case study No. : Page No. :

Computation of Regression Co-efficient :-

Regression co-efficient can be worked out from both un-replicated and replicated data. For calculation of regression co-efficient from un-replicated data three estimates, viz., (1) Sum of all obeervations on x and y ($\Sigma x, \Sigma y$) variables, (2) their sum of squares [Σx^2 and Σy^2] and (3) sum of products of all observations on x and y variables (Σxy).

Then regression co-efficient can be worked out as follows:

$$byx = \Sigma xy - (\Sigma x \cdot \Sigma y) / \Sigma y^2 - (\Sigma y)^2$$

$$bxy = \Sigma xy - (\Sigma x \cdot \Sigma y) / \Sigma x^2 - (\Sigma x)^2$$

In case of replicated data, first analysis of variances and co-variances is performed and then regression co-efficient is worked out as given below:

$$byx = cov.(xy) / vx, \text{ and } bxy = cov.(xy) / vy.$$

where, cov = co-variance between x and y

vx = variance of x.

vy = variance of y.

The significance of regression co-efficient is generally tested with the help of t-test.

First t is worked out as given below:

$$t = byx / SE(b).$$

The calculated value of t is compared with the table value of t at desired level of significance and appropriate degrees of freedom. If the calculated value of t is greater than table value, it is considered significant and vice versa.

The value of dependent variable can be predicated with the value of independent variable. By substitution the value of dependent variable we can get value of independent variable.

Applications of Regression Co-efficient in Genetics:-

Regression analysis has wide applications in the field of genetics and breeding as given below:

- * It helps in finding out a cause and effect relationship. between two or more plant characters.
- * It is useful in determining the important yield contributing characters.
- * It helps in the selection of elite genotypes by indirect selection for yield through independent characters.
- * It also helps in predicting the performance of selected plants in the next generation.

Properties of Regression co-efficient and regression lines:-

i) The regression co-efficients remain unchanged due to a shift of origin but change due to a shift of scale.

This property states that if the original pair of variables is (x,y) and if they are changed to the pair (u,v) where

$$u = \frac{xa}{p} \text{ and } v = \frac{yc}{q}$$

$$b_{yx} = \frac{q}{p} \times b_{vu}$$

$$\text{and } b_{xy} = \frac{p}{q} \times b_{uv}.$$

Subject :
Date :
Title of the test case :
Case study No. :
Page No. :

(ii) The two lines of regression intersect at the point (Mean of "x", mean of "y"), where x and y are the variables under consideration.

(iii) The co-efficient of correlation between two variables x & y in the simple geometric mean. of the two regression co-efficients. The sign of the correlation co-efficient would be the common sign of the two regression co-efficients.

This property says that if the two regression coefficients are denoted by byx and bxy then the co-efficient of correlation is given by

$$r = \pm\sqrt{byx \times bxy}$$

If both the regression co-efficients are negative, r would be negative and if both are positive, r would assume a positive value.

(iv) The two lines of regression coincide i.e., become identical when r = -1 (or) 1 (or) in other words, there is a perfect negative (or) positive correlation between the two variables under discussion.

(v) The two lines of regression are perpendicular to each other when r=0.

Co-efficient of Corelation :-

With mean :-

- * Find the mean value for the given variables x and y i.e., $\bar{x}$ & $\bar{y}$.
- * calculate the deviations of $x(x-\bar{x})$ and $y(y-\bar{y})$
- * Square the deviations of 'x' and 'y' series.
- * Multiply the single deviation in 'x' series with single deviation in 'y' series i.e., XY.
- * Apply the formula $r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}$.

1) Find the co-efficient of correlation from the data marks in accounting & marks in statistics.

Marks in Accounting :-	48	35	17	23	47
Marks in statistics :-	45	20	40	35	45

Sol:- Let us consider, Marks in accounting as 'x'.
Marks in statistics as 'y'.

x	y	X	Y	X^2	Y^2	XY
48	45	48-34 14	45-37 - 8	196	64	112
35	20	1	-17	1	289	-17
17	40	-17	3	289	9	-51
23	35	-11	-2	-121	4	22
47	45	13	8	169	64	104
$\Sigma x = 170$	$\Sigma y = 185$			$\Sigma X^2 = 776$	$\Sigma Y^2 = 430$	$\Sigma XY = 170$

Subject :	Date :

Title of the test case :

Case study No. :	Page No. :

$\therefore \bar{x} = \frac{\Sigma x}{N} = \frac{170}{5} = 34.$

$\bar{y} = \frac{\Sigma y}{N} = \frac{185}{5} = 37.$

$\therefore$ co-efficient of correlation $(r) = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$

$$r = \frac{170}{\sqrt{776 \times 430}}$$

$$r = \frac{170}{\sqrt{333680}} = \frac{170}{577.6} = 0.294$$

$$\boxed{\therefore r = 0.294.}$$

Conclusion:- Here, $r > 0$, then the correlation is said to be positive correlation and the variables are positively correlated.

$\therefore$ The range of the correlation is $1 > r > 0$.

Without Mean calculate the co-efficient of correlation:-

$$r = \frac{N\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{N\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

where N = Total no. of observations

xy = Product of series 'x' and series 'y'

x^2 = Square the variables in series 'x'.

y^2 = Square the variables in series 'y'.

1) calculate the coefficient of correlation from the following data.

x : 2 3 4 5 6

y : 7 9 10 14 15

Sol:- Here, N = 5.

x	y	x^2	y^2	xy
2	7	4	49	14
3	9	9	81	27
4	10	16	100	40
5	14	25	196	70
6	15	36	225	90
$\Sigma x = 20$	$\Sigma y = 55$	$\Sigma x^2 = 90$	$\Sigma y^2 = 651$	$\Sigma xy = 241$

$$r = \frac{N\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{N\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{5 \times 241 - 20 \times 55}{\sqrt{5 \times 90 - (20)^2} \cdot \sqrt{5 \times 651 - (55)^2}}$$

$$= \frac{1205 - 1100}{\sqrt{450 - 400} \cdot \sqrt{3255 - 3025}}$$

$$= \frac{105}{\sqrt{50} \cdot \sqrt{230}} = \frac{105}{7.07 \times 15.16} = \frac{105}{107.18}$$

$$\boxed{r = 0.97.}$$

Conclusion:- Here $r > 0$, the co-efficient of correlation is said to be positive & variables are positively correlated.

Subject : Date :

Title of the test case :

Case study No. : Page No. :

Co-efficient of correlation with the help of Assumed Mean :-

co-efficient of correlation, $r = \dfrac{N\Sigma dxdy - \Sigma dx \cdot \Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2} \cdot \sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$

where, N = Total no. of observations.

dx, dy = deviations of variables in series 'x' & 'y', i.e.,

$dx = x - A$

$dy = y - A$

where, A = Accumed mean in the given series

dxdy = product of deviation variables in series 'x' & 'y'.

dx^2, dy^2 = Squaring the deviations of series 'x' & 'y'.

1) calculate the co-efficient of correlation from the following data.

x : 2 3 4 5 6

y : 7 9 10 14 15

Sol:- A = 2 in series x

A = 7 in series y.

x	y	$dx\ (x-A)$ $x-2$	$dy\ (y-A)$ $y-7$	dx^2	dy^2	$dx \cdot dy$
2	7	0	0	0	0	0
3	9	1	2	1	4	2
4	10	2	3	4	9	6
5	14	3	7	9	49	21
6	15	4	8	16	64	32
		$\Sigma dx = 10$	$\Sigma dy = 20$	$\Sigma dx^2 = 30$	$\Sigma dy^2 = 126$	$\Sigma dx \cdot dy = 61$.

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

$\therefore \Sigma dx = 10, \Sigma dy = 20, \Sigma dx^2 = 30, \Sigma dy^2 = 126, \Sigma dxdy = 61.$

$$= \frac{5\times61 - 10\times20}{\sqrt{5\times30 - (10)^2} \cdot \sqrt{5\times126 - (20)^2}}$$

$$= \frac{305 - 200}{\sqrt{150-100} \cdot \sqrt{630-400}}$$

$$= \frac{105}{\sqrt{50} \cdot \sqrt{230}} = \frac{105}{7.07\times15.16} = \frac{105}{107.18}$$

$$\boxed{\therefore r = 0.97}$$

$$r > 0.$$

Conclusion:- Here. r>0, the co-efficient of correlation is said to be positive & the variables are positively correlated.

Rank correlation/ Spearsman Correlation:-

In 1940's charles edwards & Pearson proposed a method for the purpose of calculated the rank correlation. This is the simplest method.

$$r_k = 1 - \frac{6\Sigma d^2}{N^3 - N}$$

where N = No. of items

d = difference between ranks.

In rank correlation 3 situations are involved:

- * when the ranks are given.
- * when the ranks are not given.
- * when the ranks are equal.

When the ranks are given:

formula, $r_k = 1 - \frac{6 \Sigma d^2}{N^3 - N}$

where, d= deviation between the ranks i.e., $R_x - R_y$.

N = Total no, of observations.

This is the situation when the ranks are given by the examiner.

1) Two Ladies were asked the rank and different types of lipsticks the rank given by them are as follows:

Lipsticks:	A	B	C	D	E	F	G
Neelu (R_1):	2	1	4	3	5	7	6
Neena(R_2):	1	3	2	4	5	6	7

Calculate spearsman rank correlation co-efficient?

Sol: $N=7$.

Lipsticks	Neelu (R_1)	Neena (R_2)	$d(R_1-R_2)$	d^2
A	2	1	1	1
B	1	3	-2	4
C	4	2	2	4
D	3	4	-1	1
E	5	5	0	0
F	7	6	1	1
G	6	7	-1	1
				$\Sigma d^2 = 12$.

∴ Rank correlation, $r_K = 1 - \frac{6\Sigma d^2}{N^3-N}$

$$= 1 - \frac{6(12)}{7^3-7} = 1 - \frac{72}{343-7} = 1 - \frac{72}{336}$$

$$\Rightarrow \frac{336-72}{336} = \frac{264}{336} = 0.7857$$

$$\boxed{\therefore r_K = 0.7857}$$

∴ r>0, the rank correlation is said to be positive rank correlation.

When Ranks are not given:-

* Assign the ranks based on ascending (d) descending order.

* Give the ranks to the following variables in the series

Apply formula, Rank correlation, $r_K = 1 - \frac{6\Sigma d^2}{N^3-N}$

spearsman

1) Calculate the Sppearsman correlation from the following data.

years :	1	2	3	4	5	6	7
sales :	97.8	99.2	98.8	98.3	98.4	96.7	97.1
Prices :	73.2	85.8	78.9	75.8	77.2	87.2	83.8.

Sol:- Assign the ranks based on descending i.e., highest to lowest.

Years	sales (x)	Rx	Prices (y)	Ry	d = (Rx-Ry)	d^2
1	97.8	5	73.2	7	-2	4
2	99.2	1	85.8	2	-1	1
3	98.8	2	78.9	4	-2	4
4	98.3	4	75.8	6	-2	4
5	98.4	3	77.2	5	-2	4
6	96.7	7	87.2	1	6	36
7	97.1	6	83.8	3	3	9
						$\Sigma d^2 = 62$

∴ N = 7

Rank correlation, $r_K = 1 - \dfrac{6\Sigma d^2}{N^3 - N}$

$$= 1 - \frac{6(62)}{7^3 - 7} = 1 - \frac{372}{343 - 7} = 1 - \frac{372}{336}$$

$$= 1 - 1.107$$

$$\boxed{\therefore r_K = -0.107.}$$

Conclusion: Here, $r < 0$, then the co-efficient of rank correlation is negative rank correlation & the variables are negatively rank correlated.

When ranks are Equal:

Rank correlation, $r_K = 1 - \dfrac{6[\Sigma d^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \text{-----} + \frac{1}{12}(m^3-m)]}{N^3 - N}$

where, m = no. of items rank is repeating in both the series

d = deviation between the ranks.

1) Explain the rank correlation co-efficient between the variables x and y from the following pairs of observed values.

x :	50	55	65	50	55	60	50	65	70	75
y :	110	110	115	125	140	115	130	120	115	160.

Sol: Here ranks are not given. We will assign the ranks based Ascending Order.

x	R_x	y	R_y	$d = R_x - R_y$	d^2
50	2	110	1.5	0.5	0.25
55	4.5	110	1.5	3	9
65	7.5	115	4	3.5	12.25
50	2	125	7	-5	25
55	4.5	140	9	-4.5	20.25
60	6	115	4	2	4
50	2	130	8	-6	36
65	7.5	120	6	1.5	2.25
70	9	115	4	5	25
75	10	160	10	0	0
					$\Sigma d^2 = 134.$

50 repeating '3' times = $\frac{1+2+3}{3} = \frac{6}{3} = 2$ gives equal rank to all the places where 50 is present.

55 repeating '2' times = $\frac{4+5}{2} = \frac{9}{2} = 4.5$ gives equal rank to all the places where 55 is present.

65 repeating '2' times = $\frac{7+8}{2} = \frac{15}{2} = 7.5.$

110 repeating '2' times = $\frac{1+2}{2} = \frac{3}{2} = 1.5$

115 repeating '3' times = $\frac{3+4+5}{3} = \frac{12}{3} = 4.$

m = no. of repeating ranks

$= 3, 2, 2, 2, 3.$

$\therefore$ Rank correlation,
$$r_k = 1 - \frac{6\left[\Sigma d^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m)\right]}{N^3 - N}.$$

Subject : Date :

Title of the test case :

Case study No. : Page No. :

$$= 1 - \frac{6\left[134 + \frac{1}{12}(3^3-3) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(3^3-3)\right]}{(10)^3 - 10}$$

$$= 1 - \frac{6\left[134 + \frac{1}{12}(27-3) + \frac{1}{12}(8-2) + \frac{1}{12}(8-2) + \frac{1}{12}(8-2) + \frac{1}{12}(27-3)\right]}{1000-10}$$

$$= 1 - \frac{6\left[134 + \frac{1}{12}\times 24 + \frac{1}{12}\times 6 + \frac{1}{12}\times 6 + \frac{1}{12}\times 6 + \frac{1}{12}\times 24\right]}{990}$$

$$= 1 - \frac{6[134 + 2 + 0.5 + 0.5 + 0.5 \times 2]}{990} = 1 - \frac{6[134 + 5.5]}{990}$$

$$= 1 - \frac{6[139.5]}{990}$$

$$= 1 - \frac{837}{990}$$

$$= 1 - 0.844 \quad 1 - 0.845$$

$$\boxed{\therefore r_K = 0.155}$$

Conclusion:- Here $r > 0$, then the co-efficient of rank correlation is positive rank correlation & the variables are positively rank correlated.

Concurrent Deviation Method:-

This method is very useful and very simplest method to calculate the correlation. In this method, to identify the direction of change of 'x' variable & 'y' variable

Apply formula, $r = \pm\sqrt{\frac{\pm 2c - N}{N}}$

where, c = concurrent deviations. i.e., d = dx·dy +ve signs.

N = No. of pairs.

1) Calculate the co-efficient of concurrent deviation from the following data.

x : 60 55 50 56 30 70 40 35 80 80 75

y : 65 40 35 75 63 80 35 20 80 60 60.

Soln

x	Dx	y	Dy	d=(dx×dy)
60		65		
55	−	40	−	+
50	−	35	−	+
56	+	75	+	+
30	−	63	−	+
70	+	80	+	+
40	−	35	−	+
35	−	20	−	+
80	+	80	+	+
80	0	60	−	0
75	−	60	0	0
				c = 8

∴ No. of positive signs, c=8; N=10 (no. of pairs is 10)

$$r = \pm\sqrt{\pm\frac{2c-N}{N}}$$

$$= \pm\sqrt{\pm\frac{2(8)-10}{10}} = \pm\sqrt{\pm\frac{16-10}{10}} = \pm\sqrt{\frac{6}{10}} = \pm\sqrt{\frac{3}{5}}$$

$$= \pm\sqrt{0.6}$$

$$\boxed{r = \pm 0.7745.}$$

Case study:-4

1) 10 competitors in a beauty contest are ranked by 3 judges in the following order.

Judge-1 :	1	6	5	10	3	2	4	9	7	8
Judge-2 :	3	5	8	4	7	10	2	1	6	9
Judge-3 :	6	4	9	8	1	2	3	10	5	7

Use the rank correlation co-efficient to determine which pair of judges has the nearest approach to common tastes in beauty?

Sol:- In order to findout which pair of judges has the nearest approach to common tastes in beauty. we compare rank correlation between ~~first~~ judge & second judge & third judge.

i) Calculate rank correlation between 1st & 2nd judges.

ii) calculate rank correlation between 2nd & 3rd judges.

iii) calculate rank correlation between 3rd & 1st judges

1st judge, 2nd judge & 3rd judge are considered as

R_1, R_2, R_3.

R_1	R_2	R_3	$D_1=(R_1-R_2)$	D_1^2	$D_2(R_2-R_3)$	D_2^2	$D_3=(R_3-R_1)$	D_3^2
1	3	6	-2	4	-3	9	5	25
6	5	4	1	1	1	1	-2	4
5	8	9	-3	9	-1	1	4	16
10	4	8	6	36	-4	16	-2	4
3	7	1	-4	16	6	36	-2	4
2	10	2	-8	64	8	64	0	0
4	2	3	2	4	-1	1	-1	1
9	1	10	8	64	-9	81	1	1
7	6	5	1	1	1	1	-2	4
8	9	7	-1	1	2	4	-1	1
			$\Sigma D_1^2=200$	$\Sigma D_1^2=200$	$\Sigma D_2^2=214$	$\Sigma D_2^2=214$		$\Sigma D_3^2=60$

$$r_k = 1-\frac{6\Sigma d^2}{N^3-N}$$

$N=10.$

(i) Rank correlation between 1st & 2nd judges

$$r_{12} = 1-\frac{6\Sigma d_1^2}{N^3-N}$$

$$=1-\frac{6(200)}{10^3-10} = 1-\frac{1200}{1000-10} = 1-\frac{1200}{990} = 1-1.212$$

$$\boxed{\therefore r_{12} = -0.212}$$

(ii) Rank correlation between 2nd & 3rd judges

$$r_{23} = 1-\frac{6\Sigma d_2^2}{N^3-N} = \frac{1-6(214)}{1000-10} = 1-\frac{1284}{990}$$

Subject : Date :

Title of the test case :

Case study No. : Page No. :

$= 1 - 1.296$

$\boxed{\therefore r_{23} = -0.296.}$

(iii) Rank correlation between 3rd & 1st judges

$$r_{31} = 1 - \frac{6\Sigma d_3^2}{N^3-N} = 1 - \frac{6(60)}{10^3-10} = 1 - \frac{360}{1000-10} = 1 - \frac{360}{990}$$

$$= 1 - 0.3636$$

$\boxed{\therefore r_{31} = 0.6364.}$

Conclusion:- Using the rank correlation coefficient to 1st & 3rd judge has the nearest approach on common taste.

Regression:-

The word regression is used by the "Sir Frances Gatton" in the year 1877. Regression Analysis is the attempt to establish the ~~relation~~ relationship between two variable. In regression analysis, one variable is considered as dependent & other is independent.

Regression Equations:-

→ Regression equation x on y

$$x = a + by.$$

$$\Sigma x = Na + b\Sigma y.$$

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

Prepared By:
[redacted]
Asst. Prof.

→ Regression equation y on x

$y = a + bx$

$\Sigma y = Na + b\Sigma x$.

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

1) From the following data obtain the two regression equations.

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

Sol:

x	y	x^2	y^2	xy
6	9	36	81	54
2	11	4	121	22
10	05	100	25	50
4	8	16	64	32
8	7	64	49	56
$\Sigma x = 30$	$\Sigma y = 40$	$\Sigma x^2 = 220$	$\Sigma y^2 = 340$	$\Sigma xy = 214$

Here, $N = 5$, $\Sigma x = 30$, $\Sigma y = 40$, $\Sigma x^2 = 220$, $\Sigma y^2 = 340$, $\Sigma xy = 214$.

Regression equation x on y :

$x = a + by$

$\Sigma x = Na + b\Sigma y \rightarrow ①$

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

Substitute the values in the equations

$30 = 5a + 40b \rightarrow ①$

$214 = 40a + 340b \rightarrow ②$

Multiply the equation ① by ⑧.

$① \times 8 \Rightarrow 240 = 40a + 320b$

$214 = 40a + 340b$

$(-) \quad (-) \quad (-)$

$26 = -20b$

$-20b = 26$

$-b = \frac{26}{20}$

$-b = 1.3$

$\boxed{\therefore b = -1.3}$

Substitute $b = -1.3$ in the equation ①

$30 = 5a + 40(-1.3)$

$30 = 5a - 52$

$30 + 52 = 5a$

$82 = 5a \Rightarrow a = \frac{82}{5} \Rightarrow \boxed{a = 16.4.}$

Substitute a,b values in equation

$x = a + by$

$\boxed{\therefore x = 16.4 - 1.3y.}$

Regression equation Y on x :-

$y = a + bx$

$\Sigma y = Na + b\Sigma x \rightarrow ①$

$\Sigma xy = a\Sigma x + b\Sigma x^2 \rightarrow ②$

Substitute the values in equations ① & ②.

$$40 = 5a + b(30) \rightarrow ①$$

$$214 = a(30) + b(220) \rightarrow ②.$$

$$40 = 5a + 30b \rightarrow ①$$

$$214 = 30a + 220b \rightarrow ②$$

Multiply the equation ① by 6.

$$①\times 6 = 240 = \cancel{30a} + 180b$$

$$214 = \cancel{30a} + 220b$$

$$(-) \quad (-) \quad (-).$$

$$26 = -40b$$

$$b = -\frac{26}{40}$$

$$\boxed{\therefore b = -0.65}$$

Substitute 'b' value in equation ①

$$40 = 5a + 30(-0.65)$$

$$40 = 5a - 19.5$$

$$5a = 40 + 19.5$$

$$5a = 59.5 \Rightarrow a = \frac{59.5}{5}$$

$$\boxed{\therefore a = 11.9.}$$

Substitute $a = 11.9$, $b = -0.65$ in equation

$$y = a + bx$$

$$y = 11.9 - 0.65x$$

$$\boxed{\therefore y = 11.9 - 0.65x.}$$

Subject : Date :

Title of the test case :

Case study No. : Page No. :

With the help of Mean :-

* Regression equation x on y $\Rightarrow x-\bar{x} = \frac{\Sigma xy}{\Sigma y^2}(y-\bar{y})$

* Regression equation y on x $\Rightarrow y-\bar{y} = \frac{\Sigma xy}{\Sigma x^2}(x-\bar{x})$.

1) From the following data obtain the regression equations.

x : 6 2 10 4 8

y : 9 11 5 8 7.

Sol :-

X	y	$x=(x-\bar{x})$	$y=(y-\bar{y})$	x^2	y^2	xy
6	9	0	1	0	1	0
2	11	−4	3	16	9	−12
10	5	4	−3	16	9	−12
4	8	−2	0	4	0	0
8	7	2	−1	4	1	−2
$\Sigma x=30$	$\Sigma y=40$			$\Sigma x^2=40$	$\Sigma y^2=20$	$\Sigma xy=-26$.

Mean, $\bar{x} = \frac{\Sigma x}{N} = \frac{30}{5} = 6$.

$$\boxed{\therefore \bar{x} = 6}$$

· Mean, $\bar{y} = \frac{\Sigma y}{N} = \frac{40}{5} = 8$

$$\boxed{\therefore \bar{y} = 8.}$$

Regression equation x on y :-

x on y $\Rightarrow x-\bar{x} = \frac{\Sigma xy}{\Sigma y^2}(y-\bar{y})$

$$x-6 = \frac{-26}{20}(y-8)$$

$$x-6 = -1.3\ (y-8)$$

$$x-6 = -1.3y + 1.3(8)$$

$$x-6 = -1.3y + 10.4$$

$$x = -1.3y + 10.4 + 6$$

$$x = -1.3y + 16.4$$

$$\boxed{\therefore x = 16.4 - 1.3y.}$$

Regression equation y on x :-

$$y \text{ on } x \Rightarrow y-\bar{y} = \frac{\Sigma xy}{\Sigma x^2}\ (y-(x-\bar{x})$$

$$y-8 = \frac{-26}{40}\ (x-6)$$

$$y-8 = -0.65(x-6)$$

$$y-8 = -0.65x + 6(0.65)$$

$$y-8 = -0.65x + 3.90$$

$$y-8 = -0.65x + 3.9$$

$$y = -0.65x + 3.9 + 8$$

$$y = -0.65x + 11.9$$

$$y = 11.9 - 0.65x.$$

$$\boxed{\therefore y = 11.9 - 0.65x.}$$

Deviation | with the help of Assumed Mean :-

Regression equation x on y :-

$$(x-\bar{x}) = r \cdot \frac{\bar{x}}{y}\ (y-\bar{y})$$

$$r\ \frac{\bar{x}}{\bar{y}} = \frac{N\Sigma dxdy - \Sigma dx \cdot \Sigma dy}{N\Sigma dy^2 - (\Sigma dy)^2}.$$

Subject :
Date :
Title of the test case :
Case study No. :
Page No. :

Regression equation y on x :-

$$y-\bar{y} = \bar{r}\cdot\frac{\bar{y}}{\bar{x}}(x-\bar{x})$$

$$r\cdot\frac{\bar{y}}{\bar{x}} = \frac{N\Sigma dx\, dy - \Sigma dx\cdot\Sigma dy}{N\Sigma dx^2 - (\Sigma dx)^2}.$$

1) obtain the regression equations from the following data.

x : 6 2 10 4 8.

y : 9 11 5 8 7.

Sol:- Assumed mean in x = 2, y = 5

N = 5.

x	y	dx=(x-A)	dy=(y-A)	dx^2	dy^2	dx dy.
6	9	4	4	16	16	16
2	11	0	6	0	36	0
10	5	8	0	64	0	0
4	8	2	3	4	9	6
8	7	6	2	36	4	12
Σx=30	Σy=40	Σdx=20	Σdy=15	$\Sigma dx^2=120$	$\Sigma dy^2=65$	Σdx dy = 34

Mean, $\bar{x} = \frac{\Sigma x}{N} = \frac{30}{5} = 6.$ Mean, $\bar{y} = \frac{\Sigma y}{N} = \frac{40}{5} = 8$

$\boxed{\therefore \bar{x}=6}$

$\boxed{\therefore \bar{y}=8.}$

Regression equation x on y :-

$$(x-\bar{x}) = r\cdot\frac{\bar{x}}{\bar{y}}(y-\bar{y}).$$

$$r \cdot \frac{\bar{x}}{\bar{y}} = \frac{N\Sigma dxdy - \Sigma dx \cdot \Sigma dy}{N\Sigma dy - (\Sigma dy)^2}$$

$$= \frac{5(34) - 20 \times 15}{5(65) - (15)^2}$$

$$= \frac{170 - 300}{325 - 225} = \frac{-130}{100} = -1.3.$$

$$\boxed{\therefore r \cdot \frac{\bar{x}}{\bar{y}} = -1.3.}$$

Substitute $r \cdot \frac{\bar{x}}{\bar{y}}$ in equation x on y.

$$x - \bar{x} = r \cdot \frac{\bar{x}}{\bar{y}} (y - \bar{y})$$

$$x - 6 = -1.3(y - 8)$$

$$x - 6 = -1.3y + 1.3(8)$$

$$x - 6 = -1.3y + 10.4$$

$$x = -1.3y + 10.4 + 6$$

$$x = -1.3y + 16.4$$

$$\boxed{\therefore x = 16.4 - 1.3y.}$$

Regression equation y on x :-

$$(y - \bar{y}) = r \cdot \frac{\bar{y}}{\bar{x}} (x - \bar{x})$$

$$r \cdot \frac{\bar{y}}{\bar{x}} = \frac{N\Sigma dxdy - \Sigma dx \Sigma dy}{N\Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{5(34) - 20(15)}{5(120) - (20)^2} = \frac{170 - 300}{600 - 400} = \frac{-130}{200},$$

$$\boxed{\therefore r \cdot \frac{\bar{y}}{\bar{x}} = -0.65.}$$

Substitute $r \cdot \frac{\bar{y}}{\bar{x}}$ in regression equation y on x.

$$y-\bar{y} = \bar{r} \cdot \frac{\bar{y}}{\bar{x}} (x-\bar{x})$$

$$y-8 = -0.65 (x-6)$$

$$y-8 = -0.65x + 6(0.65)$$

$$y-8 = -0.65x+3.9$$

$$y = -0.65x+3.9+8$$

$$y = -0.65x+11.9$$

$$\boxed{\therefore y = 11.9-0.65x.}$$

Standard Error calculation with help of Regression Equations :-

* calculate the regression equations with the help of deviation with Assumed mean

* Then calculate the standard error value.

$$x \text{ on } y \Rightarrow \sqrt{\frac{\Sigma(x-x_c)^2}{N}}$$

$$y \text{ on } x \Rightarrow \sqrt{\frac{\Sigma(y-y_c)^2}{N}}$$

x_c, y_c = change values (or) standard error value of the given variables.

1) Calculate standard error with the help of regression.

x : 6 2 10 4 8

y : 9 11 5 8 7.

Solution Assumed mean in x=2, y=5.

x	y	$dx=(x-A)$	$dy=(y-A)$	dx^2	dy^2	$dxdy$.
6	9	4	4	16	16	16
02	11	0	6	~~36~~ 0	36	0
10	5	8	0	64	0	0
4	8	2	3	4	9	6
8	7	6	2	36	4	12
$\Sigma x=30$	$\Sigma y=40$	$\Sigma dx=20$	$\Sigma dy=15$	$\Sigma dx^2=120$	$\Sigma dy^2=65$	$\Sigma dxdy=34$

Mean, $\bar{x}=\frac{\Sigma x}{N}$

$=\frac{30}{5}$

$\boxed{\bar{x}=6.}$

Mean, $\bar{y}=\frac{\Sigma y}{N}$

$=\frac{40}{5}$

$\boxed{\bar{y}=8}$

Regression X on y:

$$x-\bar{x} = r.\frac{\bar{x}}{\bar{y}}(y-\bar{y})$$

$$r.\frac{\bar{x}}{\bar{y}} = \frac{N\Sigma dxdy - \Sigma dx\Sigma dy}{N\Sigma dy^2-(\Sigma dy)^2}$$

$$= \frac{5(34)-20\times15}{5\times65-(15)^2}$$

$$=\frac{170-300}{325-225}$$

$$=\frac{-130}{100}=-1.3$$

$$\boxed{r.\frac{\bar{x}}{\bar{y}}=-1.3.}$$

Subject :

Date :

Title of the test case :

Case study No. :

Page No. :

Substitute $r.\frac{\bar{x}}{\bar{y}}$ in equation x on y.

$$x-6 = -1.3(y-8)$$

$$x-6= -1.3y+1.3(8)$$

$$x-6=-1.3y+10.4$$

$$x=-1.3y+10.4+6$$

$$\boxed{\therefore x=16.4-1.3y}$$

Regression y on x :-

$$y-\bar{y} = r.\frac{\bar{y}}{\bar{x}}(x-\bar{x})$$

$$r.\frac{\bar{y}}{\bar{x}} = \frac{N\Sigma dxdy - \Sigma dx.\Sigma dy}{N\Sigma dx^2-(\Sigma dx)^2}$$

$$= \frac{5(34)-20\times 15}{5(120)-(20)^2} = \frac{170-300}{600-400} = -\frac{130}{200} = -0.65$$

$$\boxed{r.\frac{\bar{y}}{\bar{x}} = -0.65}$$

Substitute $r.\frac{\bar{y}}{\bar{x}}$ value in y on x.

$$y-8 = r.\frac{\bar{y}}{\bar{x}}(x-6)$$

$$y-8=-0.65(x-6)$$

$$y-8=-0.65x+6(0.65)$$

$$y-8=-0.65x+3.9.$$

$$y=-0.65x+3.9+8$$

$$\boxed{\therefore y=11.9-0.65x}.$$

x on y - standard errors :-

$x = 16.4 - 1.3y$ $y = 9, 11, 5, 8, 7.$

substitute y values in $x = 16.4 - 13 \cdot y.$

$9 \Rightarrow 16.4 - 1.3(9)$

$= 16.4 - 11.7.$

$x_c = 4.7.$

$11 \Rightarrow 16.4 - 1.3(11)$

$= 16.4 - 14.3$

$x_c = 2.1.$

$5 \Rightarrow 16.4 - 1.3(5)$

$= 16.4 - 6.5$

$x_c = 9.9$

$8 \Rightarrow 16.4 - 1.3(8)$

$= 16.4 - 10.4$

$x_c = 6$

$7 \Rightarrow 16.4 - 1.3(7)$

$= 16.4 - 9.1$

$x_c = 7.3$

$y = 11.9 - 0.65x$ $x = 6, 2, 10, 4, 8.$

substitute x values in $y = 11.9 - 0.65x.$

$6 \Rightarrow 11.9 - 0.65(6)$

$= 11.9 - 3.90$

$y_c = 8$

$2 \Rightarrow 11.9 - 0.65(2)$

$= 11.9 - 1.3$

$y_c = 10.6$

Subject : Date :

Title of the test case :

Case study No. : Page No. :

$10 \Rightarrow 11.9 - 0.65(10)$

$= 11.9 - 6.5$

$y_c = 5.4.$

$4 \Rightarrow 11.9 - 0.65(4)$

$= 11.9 - 2.6$

$y_c = 9.3$

$8 \Rightarrow 11.9 - 0.65(8)$

$= 11.9 - 5.20$

$y_c = 6.7.$

x	y	x_c	y_c	$(x-x_c)$	$(y-y_c)$	$(x-x_c)^2$	$(y-y_c)^2$
6	9	4.7	8	1.3	1	1.69	1.
2	11	2.1	10.6	-0.1	0.4	0.01	0.16
10	5	9.9	5.4	0.1	-0.4	0.01	0.16
4	8	6	9.3	-2	-1.3	4	1.69
8	7	7.3	6.7	0.7	0.3	0.49	0.09
						$\Sigma(x-x_c)^2 =$ 6.20	$\Sigma(y-y_c)^2 =$ 3.10.

$\Sigma(x-x_c)^2 = 6.2$

$\Sigma(y-y_c)^2 = 3.1.$

$$x \text{ on } y = \sqrt{\frac{\Sigma(x-x_c)^2}{n}} = \sqrt{\frac{6.2}{5}} = \sqrt{1.24} = 1.113.$$

$$y \text{ on } x = \sqrt{\frac{\Sigma(y-y_c)^2}{n}} = \sqrt{\frac{3.1}{5}} = \sqrt{0.62} = 0.737.$$

Standard error in x on y = 1.113.

Standard error in y on x = 0.737.

BIMK

Subject :
Title of the test case :
Case study No. :
Date :
Page No. :

UNIT - 3

PROBABILITY

Meaning and Definition of Probability :-

* The word probability is very commonly used in day-to-day conversation and generally people have no clear idea about its meaning.

* The probability of a given event is a "expression of chance of occurance of an event."

* Probability is a number which ranges from '0' to '1'.
'0' is a failure case (or) event not occur.
'1' is a success case (or) event can occur.

According to American heritage dictionary
"Probability is the branch of mathematics that studies the chance of occurance of random events in order to predict the behaviour of a defined system".

Significance of Probability in Business Applications :-

* Probability theory has been developed & employed to treat and solve many weighting problems

* Probability is the foundation of the classical decision procedures of estimation & testing.

* Probability models can be very useful for making predictions

* Probability is concerned with the construction of econometric models with managerial decisions on planning and control with the occurance of accidents, of all kinds & with random disturbances in an electrical mechanism.

* Probability is involved in the observation of the life span of a radio active atom.

→ The phenotypes of the offspring.

→ The crossing of two species of plants.

→ The discussion about sex of an unborn baby etc.,

* Probability has become an indispensable tool for all types of formal studies that involve uncertainity.

* It should be noted that the concept of probability is employed not only for various types of scientific investigations, but also for many problems in everyday life.

* The probability theory provides a media of coping up with uncertainity.

* Highlighting the importane of probability theory is a method of decisions making under uncertainity.

Note :- Formula for getting the Probability.

$$P(E) = \frac{\text{Number of favourable cases}}{\text{Total number of likely cases}}$$

$$P(E) = \frac{P(S)}{P(N)}.$$

Subject : Date :
Title of the test case :
Case study No. : Page No. :

where, $p(s)$ = favouqable cases i.e., n_{c_r}

$p(N)$ = Total no. of cases i.e., N_{c_r}.

1) A bag contains 10 black & 20 white balls, a ball is drawn at random. what is the probability that it is black?

Sol: Total no. of balls in a bag = 20 white + 10 black balls
= 30 balls.

No. of black balls = 10

No. of white balls = 20.

what is the probability of getting a black ball.

$$P(E) = \frac{10_{c_1}}{30_{c_1}}$$

$$= \frac{10}{30}$$

$$= 1/3$$

$$= 0.333$$

$$\boxed{P(E) = \frac{1}{3} (or) 0.333.}$$

what is the probability of not getting a black ball i.e.,

$$= 1 - P(E)$$

$$= 1 - 0.33$$

$$= 0.67.$$

Note:- Sum of the probability is equal to '1'.

i.e., combination of both success and failure cases.

$$p + q = 1.$$

where, $p =$ success case

$q =$ failure case.

* The probability getting a success case is 'p' is known. we can get the failure case

$$q = 1 - p.$$

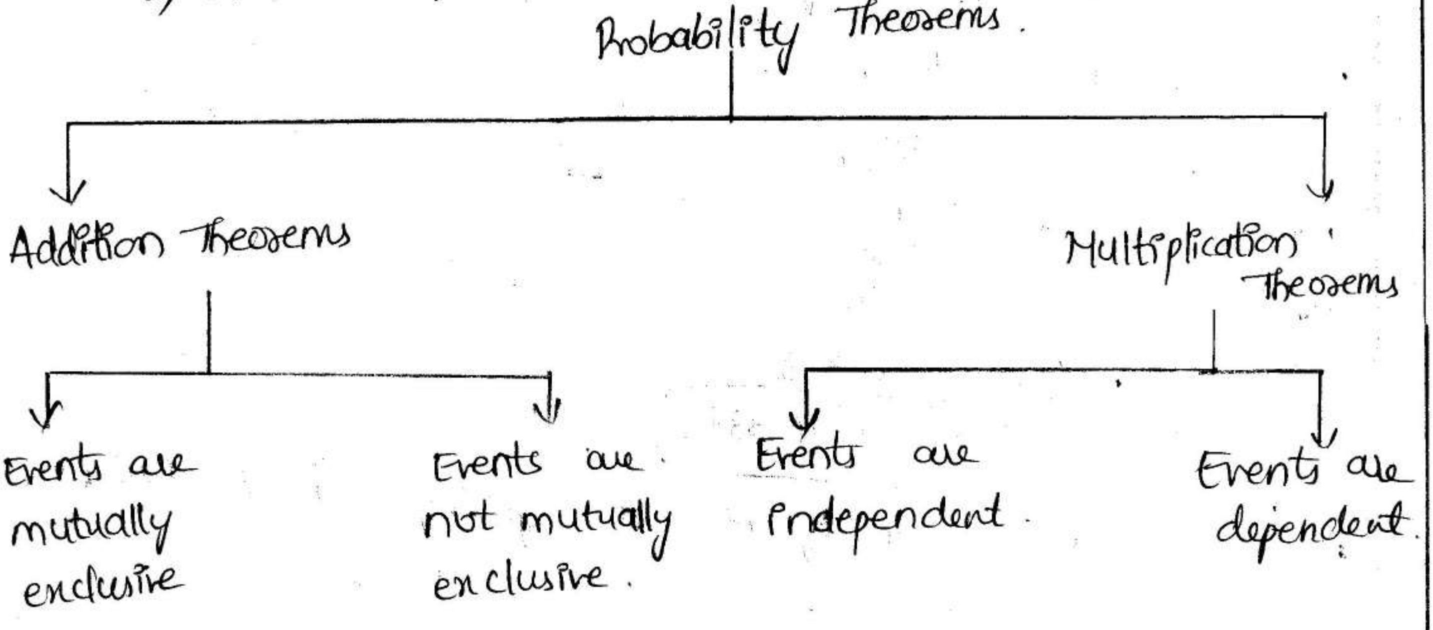
i.e., failure case is equal to difference between the sum of the probabilities to the success case.

Theories of Probability:-

There are 2 types of theories of probability namely.

1) The addition theorem.

2) The multiplication theorem.



Addition Theorems:-

This rule is related to the addition operation between two types of events to occur.

Subject	:	Date	:
Title of the test case	:		
Case study No.	:	Page No.	:

1) Mutually Exclusive Events :-

A and B are mutually exclusive the probability of the occurance of either A or B is the sum of individual probability of A & B.

Symbolically :

$$P(A \text{ or } B) = P(A) + P(B)$$

(or)

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

Proof of the Theorem :-

If an event A can happen in a_1 ways & B in a_2 ways, then the number of ways in which either event can happen is a_1+a_2. If the total no. of probabilities is n, then by definition the probability of either the first (or) the second event happening is

$$\frac{a_1+a_2}{n} = \frac{a_1}{n} + \frac{a_2}{n}$$

But, $\frac{a_1}{n} = P(A)$ & $\frac{a_2}{n} = P(B)$

Here, $P(A \text{ or } B) = P(A) + P(B)$ the theorem expand.

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C).$$

2) When events are not mutually Exclusive :-

when events are not mutually exclusive (or) in other words, it is possible for both events to occur, the addition rule must be modified.

Here, for finding the probability of one (or) more of two events that are not mutually exclusive we use the modified form of the addition theorem.

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

$P(A\cup B)$ = probability of A & B happening when A & B are not mutually exclusive.

$$P(A\cup B\cup C) = P(A) + P(B) + P(C) - P(A\cap B) - P(B\cap C) - P(A\cap C) + P(A\cap B\cap C)$$

Mutually Exclusive Events :-

1) One card is drawn from a standard pack of 52. What is the probability that it is either a king or a queen?

Sol:- There are 4 kings, 4 queens in a pack of 52 cards.

The probability that the card is drawn as a king;

i.e., $\frac{^4C_1}{^{52}C_1} = \frac{4}{52} = \frac{1}{13}$.

$P(A) = \frac{1}{13}$.

The probability that the card is drawn as a queen,

i.e., $\frac{^4C_1}{^{52}C_1} = \frac{4}{52} = \frac{1}{13}$

$P(B) = \frac{1}{13}$.

Since, the events are mutually exclusive, the probability that the card drawn is either a king (or) queen.

i.e., $P(A\cup B) = P(A) + P(B)$

$= \frac{1}{13} + \frac{1}{13} = \frac{2}{13}$

Subject : | Date :

Title of the test case :

Case study No. : | Page No. :

$\boxed{\therefore P(A \cup B) = 0.1538}$

Events are mutually not Exclusive :-

1) The managing committee of vishali welfare association formed a sub-committee of 5 persons to look into electricity problem. Profiles of 5 persons are mutually not exclusive.

1) Male age 40
2) Female age 27.
3) Male age 43.
4) Male age 65.
5) Female age 38.

If a chair person has to be selected from this what is the probability that he ~~could be~~ either female or over the 30 years?

Sol:- p(female or over 30) = p(female) + p(over 30) - p(female & over 30)

probability of female ~~p~~ p(female) $= \frac{2c_1}{5c_1} = \frac{2}{5}$.

Probability of ~~fe~~ over 30, p(over 30) $= \frac{4c_1}{5c_1} = \frac{4}{5}$

Probability of female & over 30, p(female and over 30) $= \frac{1c_1}{5c_1}$

$= \frac{1}{5}$

$\therefore$ p(female or over 30) = p(female) + p(over 30) - p(female and over 30)

$= \frac{2}{5} + \frac{4}{5} - \frac{1}{5}$

$= \frac{6}{5} - \frac{1}{5} = \frac{5}{5} = 1.$

Multiplication Theorem :- (Events are Independent).

This theorem states that if two events A & B are independent, the probability that they both will occcq is equal to the product of their individual probability.

Symbolically, if A and B are independent, then

$$p(A \cap B) = p(A) \times p(B).$$

$$p(A \cap B \cap C) = p(A) \times p(B) \times p(C).$$

Proof of the Theorem :- If an event A can happen in n_1 ways of which a_1 are successful and the events B can happen in n_2 ways of which a_2 are successful. we can combine each successful event in the first with each successful event in the second case. Thus, the total no. of successful happenings in both cases is $a_1 \times a_2$. Similarly, the total no. of possible cases is $n_1 \times n_2$.

Then by definition the probability of the occurance of both events is

$$\frac{a_1 \times a_2}{n_1 \times n_2} = \frac{a_1}{n_1} \times \frac{a_2}{n_2}$$

we know $\frac{a_1}{n_1} = p(A)$; $\frac{a_2}{n_2} = p(B)$

$$p(A \cap B) = p(A) \times p(B)$$

Events are Independent :-

1) A man wants to marry a girl having qualities.

(i) white complextion - The probability of getting such a girl is one in twenty.

(ii) Handsome dowry - The probability of getting such a person is 1 in 50.

(iii) Westernized manner - The probability of getting such a person is 1 in 100.

Find out the probability of his getting married to such a girl when the person of these 3 attributes is independent

Sol:- The probability of getting a girl with white complextion, $P(A) = \frac{1}{20}$

The probability of getting a girl with handsome dowry, $P(B) = 1/50$

The probability of getting a girl with westernized manner, $P(C) = \frac{1}{100}$

The probability of getting all qualities held in one person simultaneously i.e., $P(A \cap B \cap C)$

$$\therefore P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$$

$$= \frac{1}{20} \cdot \frac{1}{50} \cdot \frac{1}{100}$$

$$= \frac{1}{1000 \times 100}$$

$$= \frac{1}{1,00,000}$$

$$= 0.00001$$

$$\boxed{\therefore P(A \cap B \cap C) = 0.00001}$$

Conditional Probability :- (Events are dependent).

The multiplication theorem explained above is not applicable in case of dependent events. Two events A & B are said to be dependent when B can occur only when A is known to have occured. The probability attached to such an event is called the "conditional Probability" & is denoted by "P(A|B)."

If two events A and B are dependent, then the conditional probability of B given A is

$$P(B|A) = \frac{P(A\cap B)}{P(A)} \Rightarrow P(A\cap B) = P(A) \times P(B|A)$$

$$P(A|B) = \frac{P(A\cap B)}{P(B)} \Rightarrow P(A\cap B) = P(B) \times P(A|B).$$

1) A bag contains 5 white and 3 black balls. 2 balls are drawn at random one after the another without replacement. Find the probability that both balls drawn are black.

Sol:- The probability of drawing a black ball in the first attempt is $P(A) = \frac{3c_1}{8c_1}$

$$= \frac{3}{8}.$$

The probability of drawing the second ball is black. Given that the first ball is drawn black, $P(B|A) = \frac{2c_1}{7c_1} = \frac{2}{7}$.

∴ the probability that the both balls drawn the black is given by $P(A\cap B) = P(A) \cdot P(B|A)$

$$= 3/8 \times 2/7$$

Subject :
Title of the test case :
Case study No. :

Date :
Page No. :

$$= \frac{6}{56}$$

$$= 0.1071.$$

$$\boxed{\therefore P(A \cap B) = 0.1071.}$$

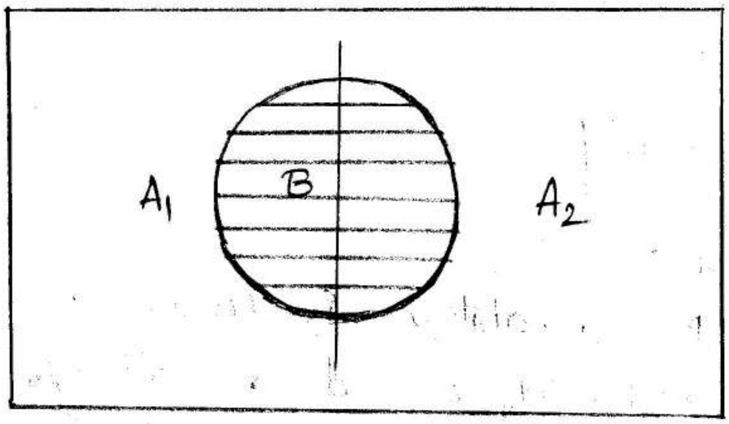
Baye's Theorem :-

The probability is known in different names, posterior probability, revised probability & inverse probability. This has been introduced by "Thomas Bayes" an english mathematician in this work known as Bayes in Decesion theory published in 1763. This theory consists of finding the probability of an event taking into account of a given sample information.

Baye's theorem is a means for qualifying uncertainity. Based on the probability theory, the theorem defines a rule for refining a hypothesis by factoring in additional evidence and back ground information and leads to a number representing the degree of probability that the hypothesis is true.

Thus a sample of 3 defective items out of 100 might be used to estimate the probability that a machine is (event A) not working propely (event B.)

It is to be denoted that the Bayesian probability is based on the formula of conditional probability where A_1 & A_2 are two events which are mutually exclusive & exhaustive & B is a simple event which intersects each of the A events as shown in the venn diagram to the right.



This is called "Posterior Probability" because it is calculated after information is taken into account. This is called "Revised Probability" as it is determined by revising the prior probabilities in the light of the additional information gathered. Further, this is called "Inverse Probability" also, as it consists of finding the probability of a problem.

However, the Bayesion (or) the posterior probabilities are always conditional probabilities which are calculated for every events as follows.

Mutually Exclusive Events :-

If an event E can only occur in combination with one of the mutually exclusive events $E_1, E_2, \text{---} E_n$ then

$$p(E_k) = \frac{[p(E_k)][p(E/E_k)]}{\sum_{i=1}^{n} p(E_i)\, p(E/E_i)} \text{ ; where } k=1,2,\text{----}n$$

Subject : Date :

Title of the test case :

Case study No. : Page No. :

Mutually Exclusive & Exhaustive Events :-

If A_1, A_2 are two mutually exclusive and exhaustive events.

$$P(A_1/B) = \frac{P(A_1)P(B/A_1)}{P(A_1) \cdot P(B/A_1) + P(A_2)\, P(B/A_2)}$$

$$P(A_2|B) = \frac{P(A_2)\, P(B/A_2)}{P(A_1)\, P(B/A_1) + P(A_2)\, P(B/A_2)}$$

1) Assume that a factory has 2 machines past records. shows that machine 1 produces 30% of the items of the output and machine 2 produces 70% of the items from the output further 5% of items produced by machine 1 were defective & only 1% produced by machine 2 were defectives. If a defective item is drawn at random. what is the probability that the defective items produced by machine 1 (or) machine 2.

Sol:- Let A_1 = items produced by machine 1.

A_2 = items produced by machine 2.

B = defective items produced by either 1 (or) 2 machines.

Probability of the items produced by machine 1

$$P(A_1) = 30\% = \frac{30}{100} = 0.3.$$

Probability of the items produced by machine 2

$$P(A_2) = 70\% = \frac{70}{100} = 0.7$$

The probability of the defective items in machine 1

$$P(B|A_1) = 5\% = \frac{5}{100} = 0.05$$

The probability of the defective items in machine 2

$$P(B|A_2) = 1\% = \frac{1}{100} = 0.01$$

Probability of the defective items produced by machine 1

$$P(A1/B) = \frac{P(A_1) \cdot P(B|A_1)}{P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2)}$$

$$= \frac{0.3 \times 0.05}{0.3 \times 0.05 + 0.7 \times 0.01}$$

$$= \frac{0.015}{0.015 + 0.007}$$

$$= \frac{0.015}{0.022}$$

$$p(A_1|B) = 0.68$$

Probability of defective items produced by machine 2

$$p(A_2|B) = \frac{p(A_2) \cdot p(B|A_2)}{p(A_1) \cdot p(B|A_1) + p(A_2) \cdot p(B|A_2)}$$

$$= \frac{0.7 \times 0.01}{0.3 \times 0.05 + 0.7 \times 0.01}$$

$$= \frac{0.007}{0.015 + 0.07}$$

$$= \frac{0.007}{0.022}$$

$$p(A_2|B) = 0.32$$

Subject : Date :

Title of the test case :

Case study No. : Page No. :

2) In a bolt factory machine A1, machine A2 and machine A3 manufactures respectively 25%, 35% & 40% of the total of their output 5,4,2 percentages are defective bolts produced by the machines. A bolt is drawn at a random from the product is found to defective. what is the probability that it was manufactured by machine 3?

Sol:

$P(A_1) = 25\% = \frac{25}{100} = 0.25$

$P(A_2) = 35\% = \frac{35}{100} = 0.35$

$P(A_3) = 40\% = \frac{40}{100} = 0.40$

The defective items produced by machine A_1 $P(B|A_1) = 5\%$

$= \frac{5}{100} = 0.05$

The defective items produced by machine A_2 $P(B|A_2) = 4\%$

$= \frac{4}{100} = 0.04$

The defective items produced by machine A_3 $P(B|A_3) = 2\%$

$= \frac{2}{100} = 0.02$

The probability of defective items by machine A_3 is

$$P(A_3|B) = \frac{P(A_3) \cdot P(B|A_3)}{P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + P(A_3) \cdot P(B|A_3)}$$

$$= \frac{0.4 \times 0.02}{0.25 \times 0.05 + 0.35 \times 0.04 + 0.40 \times 0.02}$$

$$= \frac{0.08}{0.0125+0.014+0.08}$$

$$= \frac{0.008}{0.0345}$$

$P(A_3|B) = 0.231$

Needs of Bayes Theorem :-

* The sample space is partioned into a set of mutually exclusive events ($A_1, A_2, ---- A_n$).
* With in the sample space, there exists an event B for which $p(B) > 0$.
* The analytical goal is to compute a conditional probability of the form $p(A_k|B)$.
* Atleast one of the two sets of probabilities descussed below:

(i) $P(A_k \cap B)$ for each A_k.

(ii) $P(A_k)$ and $P(B|A_k)$ for each A_k.

Features :-

* Through it deals with a conditional probability; its interepretation is different form that of the general conditional probability theorem.
* Very useful to decision making.
* The nations of priors and posterior in Bayes theorem are relative to a given sample a outcome.

Applications :-

* The theorem still prescribes multiplying the prior distribution by the likelihood function and them normalising,

to get the posteior distribution.

* As a formal theorem, Bayes theorem is valid in all common intepretations of probability.

Binomial Distribution:-

The binomial distribution also known as "Bernouli Distribution" & associated with the name of a swiss mathematician James Beenouli also known as Jacques & Jakob (1654-1705). Binomial distribution is a probability distribution, Expressing the probability of one size of dichotomous alternatives. i.e., success or failure

The distribution has been used to describe a wide variety of processes in business and the social sciences as well as other areas.

Mathematical Distribution:-

If an event E has probability of 'p' accounting in each of 'n' independent trails and that of failure in any that is $q=1-p$ then the probability that it will occur exactly 'r' times in 'n' trails is given by

$$\boxed{p(r) = n_{c_r}\, p^r\, q^{n-r}}$$

This probability distribution is called the "Binomial Probability distribution".

where, $p=$ ~~probltt~~ probability of success in a single trail.

$q = 1-P$, n = no. of trails.

r = no. of success of 'n' trails.

Obtaining Co-efficients of the Binomial :-

For obtaining co-efficients from the binomial expansion the following rules may be remembered. To find the terms of the expansion of $(q+p)^n$.

1) The first term is q^n.

2) The second term is $n_{c_1} q^{n-1} p$.

3) In each succeeding term the power of 'q' is reduced by '1' and the power of a 'p' is increased by '1'.

4) The co-efficient of any term is found by multiplying the co-efficient of the preceeding term by the power of 'q' in that preceding term and dividing the products so obtained by one more than the power of p in that preceding term. when we expand $(q+p)^n$, we get

$$(q+p)^n = q^n + n_{c_1} q^{n-1} p + n_{c_2} q^{n-2} p^2 + \cdots + n_{c_r} q^{n-r} p^r + \cdots + p^n$$

where,

$1, n_{c_1}, n_{c_2} \ldots$ are called the binomial (distribution) co-efficients.

Properties of Binomial Distribution :-

1) The shape and location of binomial distribution changes as a 'p' changes for a given 'n' (or) as 'n' changes for a given 'p'. As 'p' increases for a fixed 'n', the binomial distribution shifts to the right.

2) The mode of the binomial distribution is equal to the value of x, which has the largest probability.

3) As 'n' increases for a fixed 'p', the binomial distribution moves to the right, flattens & spreads out. the mean of the binomial distribution 'np', obviously increases as 'n' increases with 'p' held constant. For large 'n' there are more possible outcomes of a binomial expeiment and the probability associates with any particulare outcome becomes smaller.

4) ~~If~~ 'n' is the large and if neither 'p' nor 'q' is too close to zero, the binomial distribution can be closely approximated by a normal distribution with standardized variable given by $Z = \frac{x-np}{\sqrt{npq}}$. The approximation becomes better with increase 'n'.

Importance:-

The binomial probability distribution is a discrete probability distribution that useful in describing an enormous variety of real life events.

The binomial distribution can be used when:-

* The outcome of results of each trail in the process are characterised as one of two types of possible outcomes. In otherwords they are attributes.

* The possibility of outcomes of any trail does not change and is independent of the results of perious trails.

1) A fair coin is tossed thrice. find the probability of getting.

i) Exactly 2 heads..

ii) Atleast 2 heads.

Sol:- Binomial distribution.

$$p(r) = n_{C_r} \, p^r \, q^{n-r}$$

$p = \frac{1}{2}$ i.e., probability of a getting a success case.

$$q = 1 - p = 1 - \frac{1}{2} = \frac{2-1}{2} = \frac{1}{2}.$$

i) Exactly 2 heads :- $r = 2$heads

$$p(r) = n_{C_r} \, p^r \, q^{n-r}$$

$$p(2H) = 3_{C_2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{3-2}$$

$$= 3_{C_2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1$$

$$= \frac{3 \times \cancel{2}}{1 \times \cancel{2}} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1$$

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

$$p(2H) = \frac{3}{8}.$$

ii) Atleast 2 heads :- $r = (2$heads, 3heads$)$

$$p(2H) = 3_{C_2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{3-2}$$

$$= 3_{C_2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1$$

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

$$p(2H) = \frac{3}{8}$$

Subject :
Title of the test case :
Case study No. :
Date :
Page No. :

$P(3H) = {}^3C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{3-3}$

$= \frac{3\times2\times1}{1\times2\times3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^0$

$= 1\times\left(\frac{1}{2}\right)^3 \times(1)$

$P(3H) = \frac{1}{8}$.

$\therefore 2H + 3H = \frac{3}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$

2) 4 coins tossed simultaneously what is the probability of getting (i) No heads.
(ii) No ~~t~~tails :
(iii) 2 heads only (or) exactly 2 heads.

Sol:- (i) No heads :- $r=0$.

$P(0) = 4_{C0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{4-0}$

$= 1\cdot \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4$

$= 1\times1\times\frac{1}{16}$

$P(0) = 0.0625.$

(ii) No Tails :- $r=0$

$P(0) = 4_{C0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{4-0}$

$= 1\cdot \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4$

$= 1\times1\times\frac{1}{16}$

$= \frac{1}{16}$

$P(0) = 0.0625.$

(iii) 2 heads only :-

$P(2H) = 4_{C_2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{4-2}$

$= \frac{\overset{2}{\cancel{4}}\times 3}{1\times \cancel{2}} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2$

$= 6\left(\frac{1}{4}\right)\left(\frac{1}{4}\right) \Rightarrow 6\times\frac{1}{16} \Rightarrow \frac{6}{16} \Rightarrow 0.375$

Note :-

* whenever mean, standard deviation and variance are given in the binomial distribution we can consider as

mean = np.

standard deviation = $\sqrt{npq}$.

variance = $(\sqrt{npq})^2$.

1) The mean of a binomial distribution is 20 and standard deviation is 4. Find n, p & q values.

Sol:- Mean, $np = 20$

standard deviation, $\sqrt{npq} = 4$

variance, $npq = (4)^2$

$= 16$ i.e., $\sqrt{npq} = 4$ Squaring on bothsides

$(\sqrt{npq})^2 = (4)^2$

$npq = 16$

(q)

probability = $\frac{\text{variance}}{\text{mean}} = \frac{npq}{np} = \frac{16}{20} = \frac{4}{5}$ ∴ $q = \frac{4}{5}$

we know that $p+q=1$

$p = 1-q$

$= 1-\frac{4}{5} = \frac{5-4}{5} = \frac{1}{5}$

Substitute $p=\frac{1}{5}$ in mean.

$np = 20$

$n\left(\frac{1}{5}\right) = 20 \rightarrow \frac{n}{5} = 20 \Rightarrow n = 20\times 5 \Rightarrow n = 100.$

Subject :

Date :

Title of the test case :

Case study No. :

Page No. :

ancel

(Standard deviation

$\sqrt{npq} = 4$.

Squaring on bothsides) cancel.

∴ The values of n, p & q is 100, $\frac{1}{5}$, $\frac{4}{5}$.

2) The mean of a binomial distribution is 6 and variance is 4. Find n, p, q values.

Sol: Mean, $np = 6$.

variance, $npq = 4$.

$\sqrt{npq} = \sqrt{4} = 2$.

$q = \frac{\text{variance}}{\text{mean}} = \frac{npq}{np} \Rightarrow \frac{4}{6}$

$q = \frac{2}{3}$.

$p = 1-q = 1 - \frac{2}{3} = \frac{3-2}{3} = \frac{1}{3}$

Substitute $p = \frac{1}{3}$ in mean

$np = 6$

$n(\frac{1}{3}) = 6$

$\frac{n}{3} = 6 \Rightarrow n = 18$.

∴ The values of n, p & q is 18, $\frac{1}{3}$, $\frac{4}{6}$, $\frac{2}{3}$.

3) A dye is thrown 5 times if getting an even no. is a success. what is the probability of getting

(i) 4 success cases.

(ii) At least 4 success cases.

Sol: n = no. of times a dye is thrown = 5.

p = probability of getting a even no. = $\dfrac{\text{no. of items then even no. existed}}{\text{Total no. of cases}}$

$$p = \frac{3}{6}$$

$$p = \frac{1}{2}.$$

q = probability of getting a failure case

$$q = 1-p = 1-\frac{1}{2} = \frac{1}{2}.$$

$$q = \frac{1}{2}.$$

(i) 4 success cases i.e $r = 4$

$$p(r) = p(4) = {}^{5}C_{4} \left(\frac{1}{2}\right)^{4} \left(\frac{1}{2}\right)^{5-4}$$

$$= \frac{5\times\cancel{4\times3\times2}}{1\times\cancel{2\times3\times4}} \left(\frac{1}{2}\right)^{4} \left(\frac{1}{2}\right)^{1}$$

$$= 5\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right)$$

$$= 5\times\frac{1}{16}\times\frac{1}{2}$$

$$= \frac{5}{32}$$

$$p(4) = 0.156.$$

(ii) Atleast 4 success cases i.e $r \neq 0,1,2,3,4$. $r = 4,5$.

$$P(4) = {}^{5}C_{4}\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right)^{5-4}$$

$$= \frac{5\times\cancel{4\times3\times2}}{1\times\cancel{2\times3\times4}} \left(\frac{1}{16}\right)\left(\frac{1}{2}\right)$$

$$= 5\left(\frac{1}{32}\right)$$

$$= \frac{5}{32}.$$

$$p(4) = 0.156.$$

Subject : Date :

Title of the test case :

Case study No. : Page No. :

$P(5) = {}^5C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^{5-5}$

$= \frac{5\times4\times3\times2\times1}{1\times2\times3\times4\times5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^0$

$= 1\times\frac{1}{32}\times1$

$= \frac{1}{32}$

$p(5) = 0.031$

$P(4) + P(5) = 0.156 + 0.031 = 0.187$

Fitting a binomial distribution :-

When a binomial distribution is to be fitted to observe data. The following procedure is adopted.

* Determine the values of p&q. If one of these values is known the other can be found out by the simple relationship $p=1-q$ & $q=1-p$. When p&q are equal the distribution is symmetrical, for p&q may be interchanged without alternating the value of any terms & consequently terms equidistant from the two ends of the series are equal.

* Expand the binomial distribution $(q+p)^n$. The power 'n' is equal to one less than the number of terms in the expanded binomial thus when two coins are tossed $(n=2)$ there will be three terms in the binomial.

* Multiply each term of the expanded binomial by N (frequency) in order to obtain the expected frequency in each category. $\boxed{P(r) = N \times {}^nC_r\, p^r q^{n-r}.}$

1) 4 coins are tossed 160 times and the following results are obtained.

No. of heads :	0	1	2	3	4
Frequency :	17	52	54	31	6.

Fit a binomial, distribution under the assumption the coins are unbiased.

Soln: Here, $N=160$

$n=4$

$r=0,1,2,3,4$ (success cases)

$p=\frac{1}{2}$, $Q=1-P=1-\frac{1}{2}=\frac{1}{2}$.

No. of Heads	Expected frequency.
0	10
1	40
2	60
3	40
4	10

$r=0$ $p(0)= N\times n_{C_r}\, p^r q^{n-r} \Rightarrow 160\times 4_{C_0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{4-0}$

$=160\times 4_{C_0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4$

$=160\times 1\times 1\times\frac{1}{2^4}$

$=160\times\frac{1}{16}$

$p(0)=10.$

$p(1)= N\times n_{C_r}\, p^r q^{n-r}$

$=160\times 4_{C_1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{4-1}$

$=160\times 4\times\frac{1}{2}\times\left(\frac{1}{2}\right)^3$

$=\overset{40}{\cancel{160}}\times\overset{\cancel{2}\,1}{\cancel{4}}\times\frac{1}{\cancel{2}_1}\times\frac{1}{\cancel{8}\cancel{4}_1}=40.$

$p(1)=40.$

Subject :

Date :

Title of the test case :

Case study No. :

Page No. :

$P(2) = N \times {}^{n}C_{r}\, p^{r}\, q^{n-r}$

$= 160 \times 4C_2 \left(\frac{1}{2}\right)^2 . \left(\frac{1}{2}\right)^{4-2}$

$= 160 \times \frac{4\times3}{1\times2} \left(\cancel{0}\frac{1}{4}\right)\left(\frac{1}{2}\right)^2$

$> 160 \times 6 \times \frac{1}{4} \times \frac{1}{4}$

$= \overset{10}{\cancel{160}} \times \frac{6}{\cancel{16}_1}$

$P(2) = 60.$

$P(3) = 160 \times 4C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{4-3}$

$= 160 \times \frac{\cancel{4\times3\times2}}{1\times\cancel{2\times3}} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1$

$= \overset{40}{\cancel{160}} \times \overset{2\,1}{\cancel{4}} \times \frac{1}{\cancel{8}_{\cancel{4}_1}} \times \frac{1}{\cancel{2}_1}$

$P(3) = 40$

$P(4) = 160 \times 4C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{4-4}$

$= 160 \times \frac{\cancel{4\times3\times2\times1}}{\cancel{1\times2\times3\times1}} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^0$

$> 160 \times \frac{1}{24} \times \left(\frac{1}{2}\right)^0$

$= \overset{10}{\cancel{160}} \times \frac{1}{\cancel{16}_1} \times 1$

$P(4) = 10$

2) Fit a binomial distribution from the following data.

x :	0	1	2	3	4
f :	28	62	46	10	4

50%

x :	0	1	2	3	4
f :	28	62	46	10	4

fx : 0 62 92 30 16 = 200

$$\text{Mean } \bar{x} = \frac{\Sigma fx}{N} = \frac{200}{150} = \frac{4}{3}.$$

we know that, mean $np = 4/3$, but $n = 4$

$$4p = 4/3 \qquad .\, q = 1-p$$

$$p = \frac{\cancel{4} \times \cancel{4}}{3 \times \cancel{4}} \qquad q = 1 - \frac{1}{3}$$

$$p = \frac{1}{3} \qquad q = \frac{2}{3}.$$

If $r = 0$

$$p(0) = N \times n_{C_r}\, p^r\, q^{n-r}$$

$$= 150 \times 4_{C_0} \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^{4-0}$$

$$= 150 \times 1 \times 1 \times \frac{16}{81}$$

$$= 150 \times \frac{16}{81}$$

$$= \frac{2400}{81}$$

$$= 29.62.$$

$$p(1) = N \times n_{C_r}\, p^r\, q^{n-r}$$

$$= 150 \times 4_{C_1} \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^{4-1}$$

$$= 150 \times \frac{4}{1} \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^3$$

$$= 150 \times 4 \times \frac{1}{3} \times \frac{8}{27}$$

$$= 59.25.$$

$$p(2) = 150 \times 4_{C_2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^{4-2}$$

$$= 150 \times \frac{\overset{2}{\cancel{4}} \times 3}{1 \times \cancel{2}_1} \left(\frac{1}{9}\right)\left(\frac{2}{3}\right)^2$$

$$= 150 \times 6 \times \frac{1}{9} \times \frac{4}{9}$$

$$= 44.44.$$

$$P(3) = 150 \times {}^4c_3 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^{4-3}$$

$$= 150 \times \frac{4\times3\times2}{1\times2\times3} \left(\frac{1}{27}\right)\left(\frac{2}{3}\right)^1$$

$$= 150 \times \frac{\cancel{24}\,4}{\cancel{6}\,1} \times \frac{1}{27}\left(\frac{2}{3}\right)$$

$$= 150 \times 4 \times \frac{1}{27} \times \frac{2}{3}$$

$$= 150 \times 4 \times \frac{2}{27\times3}$$

$$= 14 \cdot 81.$$

$$P(4) = 150 \times {}^4c_4 \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^{4-4}$$

$$= 150 \times \cancel{\frac{4\times3\times2\times1}{1\times2\times3\times4}} \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^0$$

$$= 150 \times 1 \times \frac{1}{81} \times 1$$

$$= 1 \cdot 851.$$

Poisson Distribution :-

Poisson distribution is a discrete probability distribution and is very widely used in statistical work. It was developed by french mathematician Simeon denis poisson (1781-1840) in 1837.

Poisson distribution may be expected in cases where the chance of any individual event being a success is small. The distribution is used to describe the behaviour of rare events such as the no. of accidents on road, no. of printing mistakes in a book etc and has been called "The law of improbable events".

Mathematical Definition:-

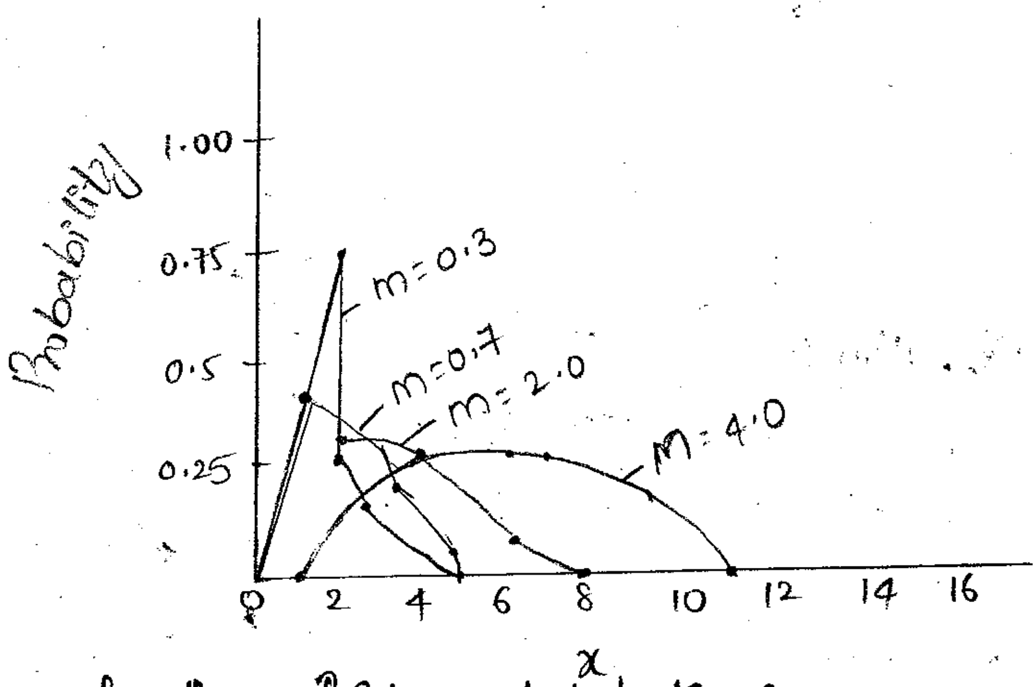
The poisson distribution $p(r)=\frac{e^{-m}m^{r}}{r!}\ m$

where, $r=0,1,2,3,4,----$

$e=2.7183$ the base of natural logarithms

m=mean of the poisson distribution.

The poisson distribution is a discrete distribution with a single parameter m. As 'm' increases the distribution shifts to the right.



Role of the Poisson distribution:-

* It is used in quality control statistics to count the no. of defects of an item.

* In biology to count the no. of bacteria.

* In physics to count the no. of practices emitted from a radio active substance.

Subject : Date :

Title of the test case :

Case study No. : Page No. :

* In Insurance problems to count the no. of cesuelities
* In waiting time problems to count the no. of incoming telephone calls (or) incoming customers.
* No. of traffic arrivals such as trucks at terminals, aeroplanes at airports ships and so fourth.
* In deteemining the no. of deaths in a distinct in a given period, say, a year, by a rare disease.
* The no. of typographicade errors per page in typed material, no. of deaths as a result of good accidents etc;
* In problems dealing with the inspection of manufactused products with the probability that any one piece is defective is very small and the lots are very large and
* To model the distribution of the no. of persons joining a queue to receive a service (or) purchase of a product.

Characters of Poison Distribution:-

Discrete distribution :- Like binomial distribution it is also a discrete probability i.e., occurances can be described by a random variable.

Main Parameter :- The main parameter is mean (m) which is equal to np i.e., $m = np$.

Form :- It is a positively skewed distributions.

No upper limit :- There is no upper limit with the no. of occurances of an event during a specified time periods

Properties :-

* The experiments results in outcomes that can be classified as successes (or) failures.
* The average no. of success (m) that occur in a specified region is known.
* The probability that a success will occur is proportional to the size of the region.
* The probability that a success will occur is an extremely small region is virtually zero.
* It is discrete probability distribution where the random variable x assumes the infinite set of values 0,1,2, ---.
* Mean = m = parameter of the distribution, variance $(\sigma)^2 = m$, S.D $(\sigma) = \sqrt{m}$, Skewness $= \frac{1}{\sqrt{m}}$ & kurtosis $= \frac{1}{m}$.
* The mode of poisson distribution is that value x which occurs with largest probability it may have either one of two models. If 'm' is not an integer, the mode is the integral value between m-1 & m. If, however m is an integer, then there are two modes which are m-1 & m.
* If x & y be two independent poisson variates with parameters m_1 & m_2 respectively, then their sum x+Y is also a poisson variate with parameter $m_1 + m_2$.
* The first, second and third new movements are respectively m, m^2+m, m^3+3m^2+m.

1) It is given that 2% of screws manufactured by a company are defective use poisson distribution to find the probability that a packet contains 100 screws.

i, No defective items (or) screws.

ii, One defective screws.

iii, Two (or) more defective screws.

Sol:- p = probability of getting the defective items = 2% $= \frac{2}{100} = 0.02$,

$$q = 1-p = 1-0.02 = 0.98$$

Mean = np, here n=100

$= 100 \times 0.02$

Mean = 2

$$p(r) = \frac{e^{-m} m^{r}}{r!}$$

i) No defective items (r=0):-

$$p(0) = \frac{e^{-2} \cdot 2^{0}}{0!} = \frac{0.135 \times 1}{1!} = 0.135.$$

ii) One defective screws:-

$$p(1) = \frac{e^{-2} \cdot 2^{1}}{1!} = \frac{0.135 \times 2}{1} = 0.270.$$

iii) Two (or) More defective items:- (r=2)

$$= 1-[(p(0)+p(1))]$$

$$= 1-[0.135+0.27] = 1-0.405 = 0.595$$

2) Suppose on an average one house in 1000 in certain district has a fire during a year if there are 2000 houses in the district. what is the probability that exactly 5 houses will have a fire during the year?

Sol:- Total no. of houses in a district, $n = 2000$.

p= probability of getting 1 house in 1000 houses in the fire accident during a year $\frac{1}{1000}$.

$$\text{mean} = np$$

$$= 2000 \times \frac{1}{1000}$$

$$\text{Mean} = 2$$

Poisson distribution $p(r) = \frac{e^{-m} m^r}{r!}$

i) Probability of getting exactly 5 houses in a fire accident during a year, $r=5$.

$$p(5) = \frac{e^{-2} (2)^5}{5!}$$

$$= \frac{0.135\ (32)}{5 \times 4 \times 3 \times 2 \times 1}$$

$$= \frac{4.32}{120}$$

$$p(5) = 0.036$$

Fitting a poison Distribution:-

The process of fitting a poison distribution is very simple. we have just obtain the value of 'm', i.e., the average occurance and calculate the frequency of '0' success. The other frequencies can be very easily calculated as follows

$N(P_0) = Ne^{-m}$

$N(P_1) = N(P_0)\times\frac{m}{1}$

$N(P_2) = N(P_1)\times\frac{m}{2}$

$N(P_3) = N(P_2)\times\frac{m}{3}$ etc.

i) The following mistakes for a page were observed in a book. No. of mistakes per page.

No. of mistakes per page (x) :	0	1	2	3	4
No. of times the mistake occur (f) :	211	90	19	5	0

Sol: Here $N = 325$ $(211+90+19+5+0)$

$fx = 0 \quad 90 \quad 38 \quad 15 \quad 0$

$\Sigma fx = 143$

Mean, $M = \frac{\Sigma fx}{N} = \frac{143}{325} = 0.44$

$e^{-m} = e^{-0.44} = 0.644.$

$NP(0) = N\times e^{-m} = 325\times 0.644$

$= 209.3.$

$NP(1) = NP(0)\times\frac{m}{1} = 209.3\times\frac{0.44}{1}$

$= 92.09$

$NP(2) = NP(1)\times\frac{m}{2} = 92.09\times\frac{\cancel{m}}{2}(0.44) = 92.09\times\frac{0.44}{2}$

$= 92.09 \times 0.22$

$= 20.25$

$NP(3) = NP(2) \times \frac{m}{3} = 20.25 \times \frac{0.44}{3} = 20.25 \times 0.146 = 2.9.$

$NP(4) = NP(3) \times \frac{m}{4} = 2.9 \times \frac{0.44}{4} = 2.9 \times 0.11$

$= 0.319.$

Assumed (or) Success cases	Expected cases.
0	209.3
1	92.09
2	20.25
3	2.9
4	0.3
	324.9

$\simeq 325.$

2) The No. of defects per unit in a Sample of 330 units of manufacturing product was found by the following.

No. of sackets :	0	1	2	3	4
No. of units :	214	92	20	3	1

Fit a poison distribution to the data under the test for goodness.

Sol:

$m = \frac{\Sigma fx}{N} = \frac{145}{330} = 0.439$

$NP(0) = N \times e^{-m} = 330 \times e^{-0.439} = 330 \times 0.6447 = 212.75.$

$NP(1) = NP(0) \times \frac{m}{1} = 212.75 \times \frac{0.439}{1} = 212.75 \times 0.439$

$= 93.39.$

$NP(2) = NP(1) \times \frac{m}{2} = \frac{0.439}{2} = 0.2195 = \underline{\frac{m}{2}}$

$= 93.39 \times \frac{m}{2} = 93.39 \times 0.2195$

$= 20.499.$

$NP(3) = NP(2) \times \frac{m}{3} = 20.499 \times \frac{0.439}{3} = 20.499 \times 0.146$

$= 2.992.$

$NP(4) = NP(3) \times \frac{m}{4} = 2.992 \times \frac{0.439}{4} = 2.992 \times 0.109$

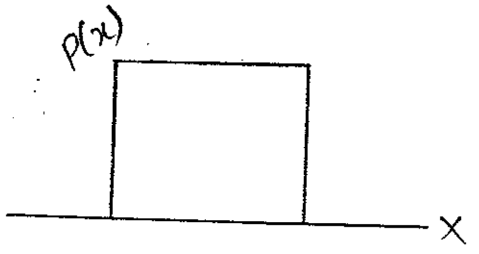
$= 0.326$

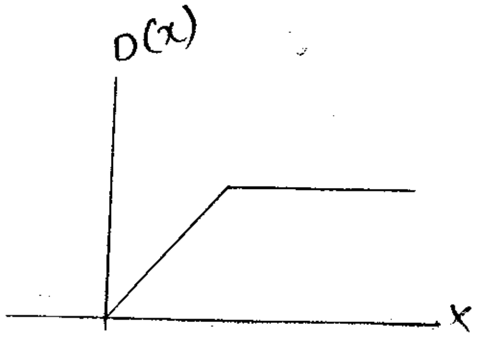
Success cases	Expected cases.
0	212.75
1	93.39
2	20.499
3	2.992
4	0.326
	329.957

≈330.

Uniform Distribution :-

A uniform distribution, sometimes also known as a rectangular distribution, is a distribution that has constant probability.





The probability density function and cummulative distribution function for a continuous uniform distribution on the interval $[a,b]$ are

$$p(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{1}{b-a} & \text{for } a \leq x \leq b \end{cases} \rightarrow ①$$

$$D(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x_2 - x_1}{b-a} & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b. \end{cases} \rightarrow ②$$

Mean and S.D of a union Distribution:-

$$\text{mean}, \mu = \frac{a+b}{2},$$

$$S.D\ \sigma = \frac{b-a}{\sqrt{12}}.$$

Probabilities in a uniform Distribution:-

The following equation is used to determining the probabilities of 'x' for a uniform distribution between a & b.

$$p(x) = \frac{x_2 - x_1}{b-a} ; \quad a \leq x_1 \leq x_2 \leq b.$$

Normal distribution:-

The normal distribution was first described by "Abraham Demoivre" as the limiting from of the binomial model in 1733. Normal distribution was rediscovered by Gauss in 1809 & by Leplace in 1812.

The normal distribution also called "the normal probability distribution".

Mathematical definition:-

The normal distribution $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)}{2\sigma^2}}$

x = value of the continuous random variable.

m = mean of the normal random variable.

e = mathematical constant approximated by 2.7183.

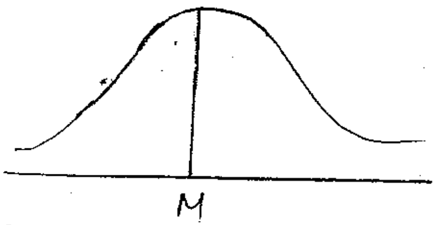
π = mathematical constant approximated by 3.1416.

$(\sqrt{2\pi} = 2.5066)$.

Graph of Normal Distribution:-

* The normal distribution can have different shapes depending on different values of M & σ but there is one & only normal distribution for any given pair of values for M & σ
* Normal distribution is a limiting case of binomial distribution when (i) $n \to \infty$

(ii) Neither p or q is very small.



* Normal Distribution is a limiting case of poison distribution when its mean m is large.
* The mean in formally distributed population lies at the centre of its normal curve.

* The two tails of the normal probability distribution extent infinitely and never too the horizontal axis.

Importance :-

* The normal distribution has the remarkable property stated in the socelled control limit theorem.
* Account to this theorem as the sample size 'n' increase the distribution of mean, $\bar{x}$ of a random sample taken from practically any population approaches a normal distribution.
* As 'n' becomes large the normal distribution saves as a good approximation of many discrete distributions.
* In theoretical statistics many problems can be solved.
* The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate.
* The normal distribution is used extensively in statistical quality control in industry in setting up of control limits.

Significance :-

* The approximate of fit a distribution of measurement under certain conditions.
* The approximate the binomial distribution and other discrete of continuous probability distributions under suitable conditions.
* The approximate the distribution of means & certain other quantities calculated from samples, especially large samples.

Properties as

* The nôrmal curve is 'bell-shaped' & symmetrical in it's appearance. If the curves were folded along it's vertical axis, the two halves would coincide.
* The height of the normal curve is at it's maximum at the mean.
* There is one maximum point of the normal curve which occurs at the mean. The height of the curve declines as we go in either direction form the mean.
* Since there is only one maximum point, the nôrmal curve is unimodel, i.e., it has only one mode.
* The points of inflection i.e., the points where the change in curvature occur are $\bar{x} \pm \sigma$
* As distinguished from binomal and passion distributed where the variable discrete the variable distributed account to the nôrmal curve is a continuous one.
* The 1st & 3rd variables are equidistant from the median.
* The mean deviation is 4th of mole preciously 0.7979 of the S.D.
* The area under the nôrmal curve distributed as follows
* Mean ±1σ covers 68.27% area - 34.135% area will lie on either side of the mean.
* Mean ±2σ covers 95.45% area.
* Mean ±3σ covers 99.73% area.

UNIT - IV

TESTING OF HYPOTHESIS

Introduction :-

The term hypothesis derives from the Greek "hypotithenai" meaning "to put under" (or) "to suppose".

Hyphothesis is a tentative conjecture explaning an observation, phenomenon, (or) scientific problem that can be tested by further observation, investigation and/(or) experimentation.

According to prof. Morris Hamburg, A hypothesis in statistics is simply a quantitative statement about population.

Statistical Hypothesis :-

A statement about population in terms of population parameter is known as a statistical hypothesis and denoted by 'H'.

Test of Hypothesis :-

A test of a hypothesis is a two action decision problem after the experimental sample values have been obtained, the two actions being the acceptance (or) rejection of the hypothesis under consideration.

Null Hypothesis :-

It is a statement which is believed to be true or it is used as a basis for argument but has not been proved it is denoted by 'H_0'.

Alternative Hypothesis :-

It is a statement of what a statistical hypothesis test is set up to establish. It is denoted by 'H_1'.

Procedure for testing of hypothesis :-

The following are various steps in testing a statistical hypothesis.

* Assume Null hypothesis : H_0

* Alternative hypothesis H_1, helps us to decide whether we have to use a single tailed (or) two tailed test.

3. Level of Significance :-

choose appropriate level of significance (α) depending on the permissible risk the α is fixed in advance before sample is drawn.

4. Test statistics :-

Compute the test statistic,

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

5. * Inference :-

We compare the computed value of z in step(4) with significant value (tabulated value) z_α = at the given level of significance 'α'.

If $|z| < z_\alpha$ we can say it is not significant i.e, the sample data do not provide us sufficient evidence against null hypothesis when may be accepted.

If $|z| > z_\alpha$, if the computed value of test statistics is more than the critical (or) significant value, then we say the null hypothesis is rejected.

Advantages :-

* Determine the focus & direction for a research effort..
* Development of a hypothesis forces the researcher to clearly state the purpose of the research activity.
* Determine what variables will not be considered in a study, as well as those that will be considered.

Disadvantages :-

* This type of tests should not be used in a mechanical fashion.
* This test do not explain the reason as to why does difference exist.
* statistical inferences based on the significance tests can't be said to be entirely correct evidences concerning the truth of the hypothesis.

Significance test for single proportion :-

Since, Sample size 'n' is large and 'x' is number of successes in 'n' independent trails with constant probability 'p' of success for each trail.

$$E(x) = np \quad \text{and} \quad v(x) = npq$$

where, $q = 1-p$.

Subject :

Date :

Title of the test case :

Case study No. :

Page No. :

It has been proved that if 'n' is large binomial distribution tends to normal distribution.

If sample size 'n' is large (i.e., $n \geq 30$) then the number of persons possessing attribute called "proportion of success".

$$p = \frac{x}{n}$$

$$\therefore E(p) = E\left(\frac{x}{n}\right)$$

$$= \frac{1}{n} \cdot E(x)$$

$$= \frac{1}{\cancel{n}} \cdot \cancel{n}P$$

$$= P$$

Thus, the sample proportion p is unbaised estimate of population proportion 'P'.

Also $V(p) = V\left(\frac{x}{n}\right)$

$$= \frac{1}{n^2} \cdot V(x)$$

$$= \frac{nPQ}{n^2}$$

$$= \frac{PQ}{n}$$

standard error $S.E(p) = \sqrt{\frac{PQ}{n}}$

then $z = \frac{p - E(p)}{S.E(p)}$

$$z = \frac{\cancel{P}p -}{\sqrt{\frac{PQ}{n}}} \; p - P$$

Note :- The limit for P at level of level of significance α are given by

$$p \pm Z\alpha\sqrt{\frac{pq}{n}}$$

1) In a sample of 1000 people in karnataka 540 are rice eaters and rest are wheat eaters can we assume both rice & wheat eaters are equally popular in this state at 1% level of significance?

Sol :- Given,

Sample $n = 1000$

Let no. of rice eaters $x = 540$.

$\therefore$ Proportion of rice eaters $p = \frac{x}{n}$

$$= \frac{540}{1000}$$

$$p = 0.54$$

Null hypothesis, (Ho) :- Both rice and wheat eaters are equally popular in the state

$$H_0 : P = 0.5$$

$$P = 0.5 \text{ and } Q = 1 - P = 1 - 0.5 = 0.5$$

Alternative hypothesis, H1 :- $P \neq 0.5$

Test statistic :- Under Ho test statistic is given by

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

$$Z = \frac{0.54 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{1000}}} = \frac{0.04}{0.0138}$$

$$Z = 2.532.$$

Significant value at 1% level for two tailed test is 2.58.

Conclusion:- Calculated value is less than significant value at 1% level of significance.

Hence, accept null hypothesis.

2) A random sample of 700 units from a large consignment showed that 200 were damaged. Find i) 95% ii) 99% confidence limits for the proportion of damaged units in the consignment.

Sol:- Given, random sample $n = 700$, $x = 200$

Proportion of damaged units $p = \frac{x}{n}$

$= \frac{200}{700}$

$p = 0.286$

$q = 1 - p = 1 - 0.286 = 0.714$

Hence, standard error $SE(p)$ is given by

$$SE(p) = \sqrt{\frac{pq}{n}}$$

$$= \sqrt{\frac{0.286 \times 0.714}{700}}$$

$$= 0.017$$

i) 95% confidence limits for p are given by

$$p \pm z_\alpha \sqrt{\frac{pq}{n}}$$

5% loss significant value is 1.96 (Z_α)

$\Rightarrow p \pm 1.96 \sqrt{\frac{pq}{n}} = 0.286 \pm 1.96 \times 0.017$

$= 0.286 \pm 0.033$

$= (0.253, 0.319)$

(ii) 99% confidence limits for p are given by $p \pm z_\alpha \sqrt{\frac{pq}{n}}$

1% loss significant value is 2.58 (Z_α)

$p \pm 2.58 \sqrt{\frac{pq}{n}} = 0.286 \pm 2.58 \times 0.017$

$= 0.286 \pm 0.044)$

$= (0.242, 0.33)$

Applications of z-test :-

* Hypothesis testing for one mean of one sample.

* Hypothesis testing for difference between means of two samples.

* Hypothesis testing for one proportion of one sample.

* Hypothesis testing. for two proportions of two samples.

* Hypothesis testing for two standard deviation of two

* Hypothesis samples.

Significance test for difference of proportions :-

Since, sample sizes n_1 and n_2 are large with x_1 and x_2 individuals possessing attributes we have

$$P_1 = \frac{x_1}{n_1}, \quad P_2 = \frac{x_2}{n_2}$$

if P_1 and P_2 are population proportions,

$E(P_1) = P_1$, $E(P_2) = P_2$

$V(P_1) = \frac{P_1 Q_1}{n_1}$, $V(P_2) = \frac{P_2 Q_2}{n_2}$

Under $H_0 \Rightarrow P_1 = P_2 = P$, $Q_1 = Q_2 = Q$. then the test statistic will becomes

$$Z = \frac{P_1 - P_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

1) Random Sample of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favour the proposal are same against that they are not at 5% loss.

Solution Given data $n_1 = 400$, $X_1 = 200$

$\Rightarrow P_1 = \frac{X_1}{n_1} = \frac{200}{400} = 0.5$

$n_2 = 600$, $X_2 = 325$

$\Rightarrow P_2 = \frac{X_2}{n_2} = \frac{325}{600} = 0.54$.

Null hypothesis; $H_0 \Rightarrow P_1 = P_2 = P$.

Assumption of null hypothesis is there is no significant difference between the opinion of men and women as per as proposal of flyover.

Alternative hypothesis, $H_1 \Rightarrow P_1 \neq P_2$.

Test statistics:-

Since samples are large, the test statistic under H_0 is $Z = \dfrac{p_1 - p_2}{\sqrt{PQ\left[1/n_1 + 1/n_2\right]}}$;

where $P = \dfrac{n_1p_1 + n_2p_2}{n_1+n_2}$

$$= \frac{400 \times 0.5 + 600 \times 0.54}{400+600}$$

$$= 0.524.$$

$Q = 1-P = 1-0.524 = 0.476.$

$$\therefore |Z| = \frac{|0.5-0.54|}{\sqrt{0.524 \times 0.476 \left(1/400 + 1/600\right)}}$$

$$|Z| = \frac{0.04}{\sqrt{0.524 \times 0.476\left(\frac{10}{2400}\right)}}$$

$$|Z| = \frac{0.04}{0.0323}$$

$$|Z| = 1.269.$$

Conclusion:-

Since $Z = 1.269$ which is less than 1.96 significant value at 5% loss.

Hence, H_0 may be accepted.

2) In a survey 800 persons out of 1000 are found tea drinkers before increase excise duty. After increase excise duty 800 persons tea drinkers out of 1200. Using standard error of proportion, state whether there is a significant

Subject :
Date :
Title of the test case :
Case study No. :
Page No. :

decrease in the consumption of tea after the increase of excise duty?

Sol:- Given data $n_1 = 1000$, $n_2 = 1200$

$X_1 = 800$, $X_2 = 800$

$P_1 = \frac{800}{1000} = 0.8$, $P_2 = \frac{800}{1200} = 0.67$.

Null hypothesis, $H_0 : P_1 = P_2$

Assume that there is no significant differsence in the consumption of tea before and after increase in excise duty.

Alternate hypothesis : $H_1 ; P_1 \neq P_2$.

Test statistics is given by under H_0 is

$$Z = \frac{P_1 - P_2}{\sqrt{PQ\left(1/n_1 + 1/n_2\right)}} \sim N(0,1)$$

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2} = \frac{16}{22}, \quad Q = 1 - P = 1 - \frac{16}{22} = \frac{6}{22}.$$

$$\therefore Z = \frac{0.8 - 0.67}{\sqrt{16/22 \times \frac{6}{22}\left(1/1000 + 1/1200\right)}}$$

$$= \frac{0.13}{0.019} = 6.842.$$

Conclusion:- Alternative 5% loss 1.96 we found evidance against Ho; Hence, we reject Ho.

Testing for means:-

In this section we will discuss the sampling of variables. For example height, weight, income, age of a group of persons.

These sampling variables each number of population provides the value of the variable.

Test of significance for single mean:-

If x_i, $i = 1, 2, 3, ---- n$ is a random sample of size 'n' from a normal population with mean 'M' and variance σ^2 then the sample mean is distributed normally with mean μ and variance $\frac{\sigma^2}{n}$; However, this result hold even in a random sampling from non-normal population provided the same size 'n' large.

Thus, for large samples, the standard normal variate corresponding to $\bar{x}$ is

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1).$$

→ In a random sampling from a large population if sampling from a finine population with size N, the correspondind limits are

$$\bar{x} \pm 1.96 \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}} \text{ and}$$

$$\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$ are 95% and 99% confidence limits.

Subject :

Title of the test case :

Case study No. :

Date :

Page No. :

1) A Sample of 400 male students is found to have a mean height of 67.47 inches can it be reasonably regarded as a sample from a large population, with mean height 67.39 inches and standard deviation 1.3 inches ($\alpha = 5\%$ loss).

Sol:- $n = 400$, $\sigma = 1.3$, $\mu = 67.3$, $\bar{x} = 67.47$

Under null hypothesis $H_0 : \mu = 67.39$.

Alternative hypothesis $H_1 : \mu > 67.39$

Test statistics is given by

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{67.47 - 67.39}{\frac{1.3}{\sqrt{400}}} = 1.23.$$

Conclusion:- We have found evidence against null hypothesis H_0. so, it can be reasonably regarded that the given sample is from the said population at 5%.

2) A random Sample of 100 articles selected from a batch of 2000 articles shows that the average diameter of the article 0.354 with standard deviation is 0.048. Find 95% confidence intervals for the average of this batch of 2000 articles?

Sol:- Given, $n = 100$, $N = 2000$, $\bar{x} = 0.354$

standard deviation = 0.048

standard error $SE(\bar{x}) = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}$

$$= \sqrt{\frac{2000-100}{2000-1}} \times \frac{0.048}{\sqrt{100}}$$

$SE(\bar{x}) = 0.00468$.

95% confidence limits for the μ are given by

$$\bar{x} \pm 1.96 \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}$$

$$= 0.354 \pm 1.96 (0.00468)$$

$$= (0.3448, 0.3632)$$

Test of significance for difference of means:-

Let $\bar{x}_1$ be the mean of a random sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and $\bar{x}_2$ be the mean of a random sample of size n_2 from a population mean μ_2 and variance σ_2^2. Then, sample sizes n_1 and n_2 are large then

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

1) In a random sample of 500, the mean is found to be 20. In another sample of 400, the mean is 15. Is the two samples drawn independently from same population with sd 4?

Sol:- $n_1 = 500$, $n_2 = 400$, $\bar{x} = 20$, $\bar{x}_2 = 15$, $\sigma = 4$

Under null hypothesis, $H_0 : \mu_1 = \mu_2$.

Alternative hypothesis, $H_1 : \mu_1 \neq \mu_2$.

Test statistic $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{1/n_1 + 1/n_2}}$

$$z = \frac{20-15}{4\sqrt{1/500 + 1/400}}$$

$$= \frac{5}{0.018}$$

$$= 277.77$$

Reject H_0.

Assumptions for students t-test :-

The following assumptions are made in the students t-test.

* The parent population from which the sample drawn is normal.
* The population, observations are independent, i.e; the given sample is random.
* The standard sample deviation is unknown.

Applications of t-distribution :-

The t-distribution has a number of applications in statistics, of which we shall discuss some of them.

* t-test for significance of single mean, population variance being unknown.
* t-test for significance of difference between two sample means, the population variances being equal

but unknown.

* t-test for significance of an observed sample correlation co-efficient.

t- test:-

The greatest contribution to the theory of small samples was made by "Sir william Sealy Gossett".

Gossett published his discovery in 1905 under the pen name 'student' and it is popularly known as t-test (or) student t-distribution (or) student's distribution.

Student's t:-

If $x_1, x_2, ---- x_n$ is a random sample of size 'n' from a normal population with mean 'u' and variance 'σ^2', the students t-statistic is defined as

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} = \frac{\bar{x} - \mu}{\sqrt{s^2/n-1}}$$

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

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

Test for Single Mean:-

1) A machine is designed to produce insulating washers for electrical devices of an average thickness of 0.025 cm. A random sample of 10 washers was found to have an average thickness of 0.024 cm with a standard deviation of 0.002 cm. Test the significance of the deviation.

Sol:- We are given, $n=10$, $\bar{x} = 0.024$ cm, $s = 0.002$ cm

$\mu = 0.025$ cm.

Subject :

Title of the test case :

Case study No. :

Date :

Page No. :

Null hypothesis :-

$H_0 : \mu = 0.025$ cm, i.e., there is no significant deviation between sample mean $\bar{x} = 0.024$ and population when $\mu = 0.025$.

Alternative hypothesis :-

$H_1 ; \mu \neq 0.025$ cm

Under Ho, the test statistics is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \rightarrow \frac{0.024 - 0.025}{0.002/\sqrt{10-1}}$$

$$\frac{-0.001 \times 3}{0.002} = -1.5.$$

Tabulated value of $t_{0.05}$ for 9 degrees of freedom = 1.833

Since, $|t| < 1.833$ is not significant between sample mean and population mean (is not significant)

2) Certain pesticide is packed into bags by a machine. A random sample of 10 bags is drawn and their contents are found to weight as follows.

50, 49, 52, 44, 45, 48, 46, 45, 49, 45.

Test if the average packing can be taken to be 50 kg.

Sol :- Null hypothesis: $H_0 = \mu = 50$ kgs.

i.e., the average packing is 50 kgs.

Alternative hypothesis : $H_1 : \mu \neq 50$ kgs.

x : 50 49 52 44 45 48 46 45 49 45.

x	$X(x-\bar{x})$	X^2
50	2.7	7.29
49	1.7	2.89
52	4.7	22.09
44	-3.3	10.89
45	-2.3	5.29
48	0.7	0.49
46	-1.3	1.69
45	-2.3	5.29
49	1.7	2.89
45	-2.3	5.29
$\Sigma x =$ 473		$\Sigma X^2 = 64.1$

Mean $= \frac{473}{10} \Rightarrow \bar{x} = 47.3$

Standard deviation $= \sqrt{\frac{\Sigma x^2}{n}}$ $(\because x - \bar{x} = X)$

Variance $(s^2) = \frac{\Sigma X^2}{n}$

$= \frac{64.1}{10}$

$s^2 = 6.41$

The test statistic is $t = \frac{\bar{x} - \mu}{\sqrt{s^2/n-1}}$

$= \frac{47.3 - 50}{\sqrt{6.41/9}}$

$= \frac{-2.7}{\sqrt{0.712}} = \frac{-2.7}{0.8438}$

$= -3.2.$

Subject :

Date :

Title of the test case :

Case study No. :

Page No. :

Tabulated value of $t_{0.05}$ for 9 degress of freedom $= 1.833$.

Since, calculated $|t|$ is greater than tabulated t, it is significant. Hence, H_0 is rejected.

2) A random samples of 10 boys had the following IQ's 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean IQ of 100. (Ans : 0.62)

Sol:- Null hypothesis : $H_0 : \mu = 100$

i.e., the assumption of a population of IQ is 100.

x	$X = (x-\bar{x})$	X^2
70	-27.2	739.84
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
88	-9.2	84.64
83	-14.2	201.64
95	-2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84
		$\Sigma X^2 = 1,833.6$

$\text{Mean} = \frac{972}{10}$

$\bar{x} = 97.2$

standard deviation $= \sqrt{\frac{\Sigma x^2}{n}}$

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

$= \frac{1833.6}{10}$

$s^2 = 183.36$

The test statistic is $t = \frac{\bar{x} - \mu}{\sqrt{s^2/n-1}}$

$= \frac{97.2 - 100}{\sqrt{183.36/9}}$

$= \frac{-2.8}{\sqrt{20.373}}$

$= \frac{-2.8}{4.513}$

$t = -0.62$

$|t| = 0.62$.

Tabulated value of $t_{0.05}$ for 9 degrees of freedom 1.833. Since, calculated $|t|$ is greater than tabulated t. It is significant. Hence, H_0 is rejected.

T-test for difference of Means:-

Suppose we want to test if two independent samples have been drawn from the two normal populations having the same means.

Let $x_1, x_2, --- x_{n_1}$ and $y_1, y_2, --- y_{n_2}$ be two independent random samples from the given normal populations.

we set up the null hypothesis $H_0 = \mu_x = \mu_y$ under the H_0 the test statistic is

$$|t| = \left| \frac{\bar{x} - \bar{y}}{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \right| \sim t_{n_1} + t_{n_2} = 2$$

where, $\bar{x} = \frac{\Sigma x}{n_1}$, $\bar{y} = \frac{\Sigma y}{n_2}$

$$s_1^2 = \frac{\Sigma(x-\bar{x})^2}{n_1}, \quad s_2^2 = \frac{\Sigma(y-\bar{y})^2}{n_2}, \quad s^2 = \frac{n_1 s^2 + n_2 s_2^2}{n_1 + n_2 - \cancel{2}\,2}$$

(follows students t-distribution with $n_1 + n_2 - 2$ dof)

1) The average number of articles produced by two machines per day are 200 and 250 with standard deviations 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machine equally efficient at 5% level of significance.

50|34

In the usual notations we are given

$n_1 = n_2 = 25$, $\bar{x} = 200$, $\bar{y} = 250$, $s_1 = 20$, $s_2 = 25$

Null hypothes-is $H_0 = \mu_1 = \mu_2$ i.e. both the machines are equally efficient.

Alternative hypothesis: $H_1 : \mu_1 \neq \mu_2$

Under the H_0 the test by statistics is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{x_1} + \frac{1}{x_2}\right)}} \quad \text{where} \quad s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

$$t = \frac{200-250}{\sqrt{533.85\left(\frac{1}{25}+\frac{1}{25}\right)}}$$

$$s^2 = \frac{25\times 400 + 25\times 625}{25+25-2}$$

$$s^2 = \frac{25625}{48}$$

$$\therefore s^2 = 533.85$$

$$= \frac{-50}{\sqrt{533.85\times 0.08}}$$

$$= \frac{-50}{\sqrt{42.708}} = \frac{-50}{6.535} = -7.65$$

Tabulated $t_{0.05}$ value for 48 = 1.67. Since, calculated $|t| >$ tabulated t, it is highly significant. Hence, H_0 is rejected and we conclude that both the machines are not equally efficient at 5% level of significance.

2) The Means of 2 random samples of size 9 and 7 are 196.42 and 198.82 respectively. The sum of the squares of the deviations from the mean are 26.94 & 18.73 respectively. can the samples be considered to have been drawn from the same normal population?

Soln: In the usual notations we are given

$n_1 = 9$, $\bar{x} = 196.42$, $\Sigma(x-\bar{x})^2 = 26.94$.

$n_2 = 7$, $\bar{y} = 198.82$, $\Sigma(y-\bar{y})^2 = 18.73$

Null hypothesis:- The samples have been drawn from the same normal populations.

i.e., $H_0 = \mu_1 = \mu_2$.

Alternative hypothesis:- $H_1 : \mu_1 \neq \mu_2$

under the H_0, the test statistics is.

$$t = \frac{\bar{x}-\bar{y}}{\sqrt{s^2\left(\frac{1}{x_1}+\frac{1}{x_2}\right)}}$$

we have $s^2 = \dfrac{\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2}{n_1+n_2-2}$

$$t = \frac{196.42 - 198.82}{\sqrt{3.26\left(\frac{1}{9}+\frac{1}{7}\right)}}$$

$$s^2 = \frac{26.94 + 18.73}{9+7-2}$$

$$= \frac{45.67}{14}$$

$$= 3.26.$$

$$= \frac{-2.40}{\sqrt{3.26 \times 0.254}}$$

$$= \frac{-2.40}{\sqrt{0.828}}$$

$$= \frac{-2.40}{0.9099} = -2.64$$

Tabulated $t_{0.05}$ for 14 dof is 1.761

Since, calculated $|t|$ is greater than tabulated t, it is significant. Hence, H_0 is rejected

3) Two different types of drugs A and B were tried on certain patients for increasing weight, 5 persons were given drug A and 7 persons were given drug B. The increase in weight in pounds are given below.

Drug A :	8	12	13	9	3	—	—
Drug B :	10	8	12	15	6	8	11

Do the two drugs differ significantly with regard to their effect in increasing weight.

(Ans:- 0.501)

F-distribution (F-test)

F-distribution was introduced by G.W. Snedecor.

The f-test is named in honour of the great

Statisfaction R.A Fisher.

f-Test for two sample standard deviations :-

Let $x_1, x_2, ---- x_n$ be a random sample of size n_1 from the first population with variance σ_1^2 and $y_1, y_2, ----y_n$ be a random sample of size n_2 from the second nomal population with variance σ_2^2. Obviously the two samples are independent.

we set up the null hypothesis as

$$H_0 = \sigma_1^2 = \sigma_2^2 = \sigma^2$$

i.e., population variances are same.

under H_0, the test statistic is

$$F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1) \text{ where } s_1^2 = \frac{\Sigma(x-\bar{x})^2}{n_1 - 1}$$

$$s_2^2 = \frac{\Sigma(y-\bar{y})^2}{n_2 - 1}$$

follows F-distribution with $(n_1 - 1, n_2 - 1)$ d.f.

Assumption in F-test :-

The F-test is based on the following assumptions.

* **Normality :-** Values in each group are nomally distributed.

* **Homogenity :-** The variance with in each group should be equal for all groups $(\sigma_1^2 = \sigma_2^2 = ---- = \sigma_k^2)$

* **Independence of Error :-** It states that the error should be independent for each value.

Applications of F-test :-

* F-test for testing the significance of an observed sample

multiple correlation.

* F-test for testing the significance of an observed sample correlation ratio.
* F-test for testing the linearity of regression.
* F-test for testing the equality of several population means, i.e., for testing $H_0 = \mu_1 = \mu_2 ---- = \mu_k$ for k normal populations.

1) Time taken by workers in performing a job by method 1 and method 2 is given below.

Method - 1 - 20 16 26 27 23 32.

Method - 2 - 27 33 42 35 32 34 38.

Do the data show that the variance of time distribution from population. From which these samples are drawn do not differ significantly?

Sol: We set up null hypothesis as $H_0 : \sigma_1^2 = \sigma_2^2$

i.e., there is no significant difference between the variances of the time distribution by the workers in performing a job by method I and method II.

Method - I

x	$x-\bar{x}$	$(x-\bar{x})^2$
20	-2.3	5.29
16	-6.3	39.69
26	3.7	13.69
27	4.7	22.09
23	0.7	0.49
22	-0.3	0.09
$\Sigma x =$ 134		$\Sigma(x-\bar{x})^2 =$ 81.34

Method - II

y	$(y-\bar{y})$	$(y-\bar{y})^2$
27	-7.4	54.76
33	-1.4	1.96
42	7.6	57.76
35	0.6	0.36
32	-2.4	5.76
34	-0.4	0.16
38	3.6	12.96
$\Sigma y =$ 241		$\Sigma(y-\bar{y})^2 =$ 133.72

$$\bar{x} = \frac{\Sigma x}{n} = \frac{134}{6} = 22.3$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{241}{7} = 34.4$$

$$S_1^2 = \frac{\Sigma(x-\bar{x})^2}{n_1-1} = \frac{81.34}{6-1} = \frac{81.34}{5} = 16.26$$

$$S_2^2 = \frac{\Sigma(y-\bar{y})^2}{n_2-1} = \frac{133.72}{7-1} = \frac{133.72}{6} = 22.28$$

Since, $S_2^2 > S_1^2$, under Ho; the test statistic is

$$F = \frac{S_2^2}{S_1^2} \sim F(n_2-1, n_1-1)$$

$$F = \frac{22.28}{16.26} = 1.37$$

Tabulated $F_{0.05}(6,5) = 4.95$.

Since, calculated F is less than tabulated F, it is not significant. Hence, Ho may be accepted at 5% level of significance.

2) It is known that the mean diameters of rivets produced by 2 firms A and B practically the same but the standard deviations may differ. For 22 rivets produced by firm A, the standard deviation is 2.9mm while for 16 rivets manufactured by firm B, the standard deviation is 3.8mm. Compute the statistic you would use to test whether the products of firm A have the same variability as those of firm B and test its significance.

Sol: Given data.

$n_1 = 22 \qquad n_2 = 16$

$s_1 = 2.9mm \ , \ s_2 = 3.8mm$

we set up the null hypothesis as $H_0 : \sigma_1^2 \geq \sigma_2^2$ i.e., the products of both the firms A and firms B have the some variability.

we have $$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{22 \times (2.9)^2}{22-1} = \frac{22 \times 8.41}{21} = \frac{185.02}{21} = 8.810$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{16 \times (3.8)^2}{16-1} = \frac{16 \times 14.44}{15} = \frac{231.04}{15} = 15.402$$

Since, $s_2^2 > s_1^2$ under Ho the test statistic is

$$F = \frac{s_2^2}{s_1^2} = \frac{15.402}{8.810} = 1.748.$$

which follows F distribution with (15,21)

Tabulated $F_{0.05}(15,21) = 2.20$.

Since, calculated F is less than the tabulated F, it is not significant at 5% level of significance. Hence, Ho is accepted.

* Design of Experiments :-

An expeimental design is a plan and a structure to test hypothesis is which the recorder either controls & manipulates one (or) more variables, it contains independent and dependent variables.

Independent Variables :-

Work shift, gender of employee, region type of machine, quality of tire.

Dependent Variable :-

A Dependent variable is the response to the different levels of the independent variables.

Principles of Experimental Design :-

* Comparison * Randomization, * Blocking.

* Replication. * Factorial Experiments.

Procedure in effective design of Experiment :-

1) Select problem.

2) Determining dependent variables

3) Determining independent vaiiables.
4) Deteimining number of levels of independent vaiiables.
5) Determining possible contributions.
6) Determining number of observations.
7) Randomization.
8) Meet ethical and legal requirements.
9) Mathematical model.
10) Data collection.
11) Data reduction.
12) Data verification.

Analysis of vaiiance (ANOVA) :-

* Analysis of vaiiance was developed by R.A. fishner.
* Analysis of vaiiance, the significance of the difference between the means of two samples can be judged through either Z- test (or) t-test, but the difficulty arises when we used ANOVA.
* ANOVA is useful in the fields of Economics, biology, education, psychology, socialogy, and business and in research of several other deciplines.
* ANOVA is essentially a procedure for testing the difference among different groups of data for homogeneity.
* ANOVA is a method of analysing the vaiiance to which a response is subject into it's various components corresponding.

to various sources of variation.

Assumptions of ANOVA :-

- * It is assumed that the universe from which the different samples are drawn for study is normally distributed.
- * It is assumed that there is no significant difference amongst the variances of the different universes from which the samples have been drawn.
- * It start with null hypothesis that $V_1 = V_2 = V_3 ------ V_n$.
- * It is assumed that the critical values of the variance ratio (F). is estimated at different levels of significance, Ex:- 5% (or) 1% etc.

Applications of ANOVA :-

- * We can explain various varieties of seeds of fertilizers (or) soils differ significantly so that a policy decision could be taken with help of 'ANOVA'.
- * various types of drugs manufactured for curing a specific disease may be studied and judged.
- * A manager of a big concern can analyze the performance of various, sales man.

Analysis of variance for one-way classification :-

Under the one way ANOVA, we consider only one factor. we determine if there are differences with in that factor.

The technique involves the following steps.

*. Calculate sum of normal, squares of the individual variables.

* calculate the sum of individual sum of the variables.

$$T = \Sigma x_1 + \Sigma x_2 + ---- + \Sigma x_n.$$

* calculate the value of connection factor $\left(\frac{T^2}{N}\right)$.

where, N = Total no. of variables.

* calculate the value of $SST = \Sigma x_1^2 + \Sigma x_2^2 + ---- + \Sigma x_n^2 - \frac{T^2}{N}$

(sum of squares for variance of total).

*. Calculate the value of $SSB = \frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n} + ---- + \frac{(\Sigma x_n)^2}{n} - \frac{T^2}{N}$

(sum of squares for variance between the samples].

* Find out the value of $SSW = SST - SSB$.

(sum of squares for variance between with in the samples]

* Draw the ANOVA Table.

* Finally, F-ratio may be worked out as,

$$F\text{-ratio} = \frac{MSB}{MSW}.$$

MSB = Mean square between samples.

MSW = Mean square with in samples.

1) 4 Machines A, B, C, D are used to produce a certain kind of cotton fabrics. samples of size 4 with each unit as. 100 square meters are selected from the outputs of the machines at random, and the number of flows in each 100

square meters are counted, with the following result.

A	B	C	D
8	6	14	20
9	8	12	22
11	10	18	25
12	4	9	23.

Do you think that there is a significant difference in the performance of the four machines.

Sol: Let us take the null hypothesis that the machines do not differ significantly in performance,

i.e., $H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4$

x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	x_4	x_4^2
8	64	6	36	14	196	20	400
9	81	8	64	12	144	22	484
11	121	10	100	18	324	25	625
12	144	4	16	9	81	23	529
$\Sigma x_1 = 40$	$\Sigma x_1^2 = 410$	$\Sigma x_2 = 28$	$\Sigma x_2^2 = 216$	$\Sigma x_3 = 53$	$\Sigma x_3^2 = 745$	$\Sigma x_4 = 90$	$\Sigma x_4^2 = 2038$

$$T = \Sigma x_1 + \Sigma x_2 + \Sigma x_3 + \Sigma x_4$$

$$= 40 + 28 + 53 + 90$$

$$T = 211.$$

correction factor $= \frac{T^2}{N} = \frac{(211)^2}{16} = \frac{44521}{16} = 2782.56.$

Sum of squares for variance of Total (SST) =

$$\Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2 - \frac{T^2}{N}$$

$= 410+216+745+2036-2782.56$:

$= 3409-2782.56$

$= 626.44$.

Sum of squares for variance between the sample (SSB)=

$$=\frac{(\Sigma x_1)^2}{n_1}+\frac{(\Sigma x_2)^2}{n_2}+\frac{(\Sigma x_3)^2}{n_3}+\frac{(\Sigma(x_4)^2}{n_4}-\frac{T^2}{N}$$

$$=\frac{(40)^2}{4}+\frac{(28)^2}{4}+\frac{(53)^2}{4}+\frac{(90)^2}{4}-2782.56$$

$$=\frac{1600}{4}+\frac{784}{4}+\frac{2809}{4}+\frac{8100}{4}-2782.56$$

$$=400+196+702.25+202.5-2782.56.$$

$$=540.69$$

Sum of squares for variance with in the samples (SSW)=

$SST-SSB$

$= 626.44 - 540.69$

$= 85.75$

N = Total no. of variables (or) sample values.

K = Number of variables types (sample (or) variables)

ANOVA Table.

① Source of Variation	② Sum of Squares	③ Degree of freedom	④ = 2/3 Mean Square
Between samples	540.69	3(K−1)	180.23
within samples	85.75	12(N−K)	7.15.

$$F-\text{ratio} = \frac{MSB}{MSW} = \frac{180.23}{7.15} = 25.207.$$

The table value for $F_{(3,12)}$ at 1% level of significance is 5.95. The calculated value of 'F' is greater than the table value - Hence, we reject the null hypothesis and conclude that there is a significant difference in the performance of the four machines.

2) A random sample is selected from each of 3 makes of rope and their breaking strength are measured, with the following results:

x_1	x_2	x_3
70	100	60
72	110	65
75	108	57
80	112	84
83	113	87
–	120	73
–	107	–

Test, whether the breaking strength of the ropes differ significantly.

Analysis of Variance for two-way classification:-

The way ANOVA techniques is used when the data are classified on the basis of two factors.

For example:- The agriculture output may be classified on the basis of different varieties of seeds and also on the basis of different varieties of fertilizers used.

In this 2 way classification 2 cases are existed.

* ANOVA technique is context of 2-way design when

Subject : Date :

Title of the test case :

Case study No. : Page No. :

repeated values are not there.

ANOVA is context of 2-way design when repeated values are there.

The following steps are involved.

* Use the coding device.
* calculate the sum of normal squares of the individual variables.
* Calculate the sum of individual sum of the variables

$$T = \Sigma x_1 + \Sigma x_2 + \text{---} + \Sigma x_n$$

* calculate the value of correction factor $\left(\frac{T^2}{N}\right)$.

where, N = Total no of variables.

* calculate the value of $SST = \Sigma x_1^2 + \Sigma x_2^2 + \text{---} + \Sigma x_n^2 - \frac{T^2}{N}$

* [calculate the value of $SSB = \frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n_2} + \text{---} + \frac{(\Sigma x_n)^2}{n} - \frac{T^2}{N}$

* Find out the value of $SSW = SST - SSB$]
* Take the total of different columns and they obtain the square of each column total and divide such squared values of each column by the number of items in the concerning column and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of square of deviations for variances between columns (SSC).
* Calculate SSR value.

* Findout the value of sum of squares of deviations for residual (or) error variance [SSE] = SST-(SSC+SSR).

* Draw the ANOVA Table.

* Test statistic $F = \frac{\text{MSB (columns)}}{\text{MSR}}, \frac{\text{MSB (rows)}}{\text{MSR}}$

MSR = Mean square residual.

1) The following table gives the number of refrigrators sold by 4. salesman in 3 months may, June & July.

	Salesman			
Month	A	B	C	D
May	50	40	48	39
June	46	48	50	45
July	39	44	40	39.

Is there a significant difference in the sales made by the 4 salesman? Is there a significant difference in the sales made during different months?

Solution: Let us take the null hypothesis that there is no significant difference between sales made by the four salesmen during different months.

∴ The given data are coded by substracting 40 from each observation calculations for a 2-criterion month & sales man.

Subject :

Date :

Title of the test case :

Case study No. :

Page No. :

Sales man

Month	x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	x_4	x_4^2	Row sum
May	10	100	0	0	8	64	-1	1	17
June	6	36	8	64	10	100	5	25	29
July	-1	1	4	16	0	0	-1	1	2
	$\Sigma x_1 =$ 15	$\Sigma x_1^2 =$ 137	$\Sigma x_2 =$ 12	$\Sigma x_2^2 =$ 80	$\Sigma x_3 =$ 18	$\Sigma x_3^2 =$ 164	$\Sigma x_4 = 3$	$\Sigma x_4^2 =$ 27	48

T = sum of all observations = 48

correction factor $= \frac{T^2}{N} = \frac{(48)^2}{12} = 192$

SSC = sum of squares between sales men (columns)

$$= \frac{(15)^2}{3} + \frac{(12)^2}{3} + \frac{(18)^2}{3} + \frac{(3)^2}{3} - 192$$

$$= (75 + 48 + 108 + 3) - 192 = 42.$$

SSR = sum of squares between months (rows)

$$= \frac{(17)^2}{4} + \frac{(29)^2}{4} + \frac{(2)^2}{4} - 192$$

$$= (72.25 + 210.25 + 1) - 192$$

$$= 91.5.$$

$SST = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2 - C.F$

$$= 137 + 80 + 164 + 27 - 192$$

$$= 216.$$

$SSE = SST - (SSC + SSR) = 216 - (42 + 91.5) = 82.5.$

Degrees of freedom $c-1 = 4-1 = 3$

$r-1 = 3-1 = 2$

$(c-1)(r-1) = 3\times2 = 6$.

The ANOVA Table :-

S.No	Sources of variation	Sum of squares	d.f	Mean squares	Variance ratio.
1.	Between salesmen	42	3	14	$F=\frac{14}{13.75} = 1.018$
2.	Between months.	91.5	2	45.75	$F=\frac{45.75}{13.75} = 3.327$
3.	Residual Error	82.5	6	13.75	$F=\frac{13.75}{13.75} = 1.00$

Conclusion :-

*1 The table value of $F = 4.75$ for $df_1 = 3$, $df_2 = 6$ & $\overset{\alpha}{d} = 0.05$, since, the calculated $F = 1.018$ is less than table value, the null hypothesis is accepted.

*2 The table value of $F = 5.14$ for $df_1 = 2$, $df_2 = 6$ and $\alpha = 0.05$, since, the calculated value of $F = 3.327$ is less than is table value, the null hypothesis is accepted.

* Perform ANOVA and decide whether the mean productivity is same (or) differs among workers.

Subject : Date :

Title of the test case :

Case study No. : Page No. :

Machine Gype :

Workers	A	B	C	D
1	40	36	48	38
2	52	44	52	42
3	35	38	45	36
4	48	32	45	34
5	40	40	50	40

Test significance levels at 5%

UNIT - V

NON-PARAMETRIC METHODS

Non-Parametric Methods :-

Practical data to estimate the parameters such as mean, variance etc and use the standard tests, they are known as "Parametric tests".

The practical data may be non-normal (or) it may me not possible to estimate the parameters of the data. The test which are used for such situations are called "Non-Parametric tests."

χ^2 - test (chy-square test) :-

The χ^2 test was first used by "karl pearson" in the year 1900. The χ^2 describes the magnitude of the descrepency between theory & observation.

χ^2-square distribution :-

The square of a standard normal variate is called a "chy square variate with 1 degree of freedom" (dof). Thus if x is a random variable following normal distribution with mean 'u' and standard deviation 'σ' then $\left(\frac{x-u}{\sigma}\right)$ is a standard normal variate.

$\therefore \left(\frac{x-u}{\sigma}\right)^2$ is a chy-square variate with 1 degree of freedom (dof).

Importance of Non Parametric Method

In statistics, nonparametric tests are methods of statistical analysis that do not require a distribution to meet the required assumptions to be analyzed (especially if the data is not normally distributed). Due to this reason, they are sometimes referred to as distribution-free tests.

Nonparametric tests serve as an alternative to parametric tests such as T-test or ANOVA that can be employed only if the underlying data satisfies certain criteria and assumptions.

Some of the Nonparametric tests:

1) Mann-Whitney U Test.
2) Wilcoxon Signed Rank Test.
3) Kruskal - Wallis Test.
4) Chi-Squared Test.

Reasons to Use Nonparametric Tests:

- The underlying data do not meet the assumptions about the population sample.
- The population sample size is too small
- The analyzed data is ordinal or nominal

Types of Tests

Nonparametric tests include numerous methods and models. Below are the most common tests and their corresponding parametric counterparts

1. Mann-Whitney U Test

The Mann-Whitney U Test is a nonparametric version of the independent samples t-test. The test primarily deals with two independent samples that contain ordinal data.

2. Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank Test is a nonparametric counterpart of the paired samples t-test. The test compares two dependent samples with ordinal data.

3. The Kruskal-Wallis Test

The Kruskal-Wallis Test is a nonparametric alternative to the one-way ANOVA. It is used to compare more than two independent groups with ordinal data.

Properties	Parametric	Non-parametric
Assumptions	Parametric statistics are based on assumptions about the distribution of population from which the sample was taken.	Nonparametric statistics are not based on assumptions, that is, the data can be collected from a sample that does not follow a specific distribution.
central tendency Value	If the **mean** more accurately represents the center of the distribution of your data, and your sample size is large enough, use a parametric test.	If the **median** more accurately represents the center of the distribution of your data, use a nonparametric test even if you have a large sample size.
Correlation	The most frequent parametric test to examine for strength of association between two variables is a **Pearson correlation (r).**	Spearman's Rho is a **non-parametric** test used to measure the strength of association between two variables, where the value r = 1 means a perfect positive correlation and the value r = -1 means a perfect negataive correlation.
Probabilistic distribution	Parametric tests assume a normal distribution of values, or **a "bell-shaped curve**."	**Non-parametric tests are** valid for both non-Normally distributed data and Normally distributed data
Population knowledge	In parametric population knowledge should be required.	In non parametric population knowledge does not required.
Used for	In parametric test that the variables in the population are measured based on an **interval scale.**	Nonparametric statistics are used when our data are measured on a **nominal or ordinal scale** of measurement.
Examples	z-test, t-test,f-test & ANOVA test.	1-sample sign test, Wilcoxon Signed Rank test, Kruskal-Wallis test, Mann-Whitney test, Spearman Rank Correlation etc..

Goodness-of-Fit

The term goodness-of-fit refers to a statistical test that determines how well sample data fits a distribution from a population with a normal distribution.

Goodness-of-fit establishes the discrepancy between the observed values and those expected of the model in a normal distribution case. There are multiple methods to determine goodness-of-fit, including the chi-square.

Goodness-of-fit tests are statistical methods that make inferences about observed values. For instance, you can determine whether a sample group is truly representative of the entire population. As such, they determine how actual values are related to the predicted values in a model. When used in decision-making, goodness-of-fit tests make it easier to predict trends and patterns in the future.

As noted above, there are several types of goodness-of-fit tests. They include the chi-square test, which is the most common, as well as the Kolmogorov-Smirnov test, and the Shapiro-Wilk test. The tests are normally conducted using computer software. But statisticians can do these tests using formulas that are tailored to the specific type of test.

Key points:

1) A goodness-of-fit is a statistical test that tries to determine whether a set of observed values match those expected under the applicable model.
2) They can show you whether your sample data fit an expected set of data from a population with normal distribution.
3) There are multiple types of goodness-of-fit tests, but the most common is the chi-square test.
4) The chi-square test determines if a relationship exists between categorical data.
5) The Kolmogorov-Smirnov test determines whether a sample comes from a specific distribution of a population.

Applications of χ^2-Test:-

chy-square distribution has a number of applications some of which are equmerated below:

1) chy-square test of goodness of fit.

2) chy-square test for independence of attributes.

3) To test if the population has a specified value of the variable σ^2.

Conditions for applying χ^2-Test:-

* N, the total number of frequencies should be reasonably large, say greater than 50

* The sample observations should be independent.

* No theoretical cell frequency should be small.

* The given distribution should not be replaced by relative frequencies of proportions but data should be given in orginal units.

Chy-square test for single sample standard deviation:-

Suppose we want to test if the given normal population has a specified variance.

$\sigma^2 = \sigma_0^2$ (say) or not

σ_0^2 = specified value.

if $x_1, x_2, \ldots x_n$ is a random sample of size 'n' from the given population.

we set up null hypothesis as $H_0 = \sigma^2 = \sigma_0^2$

under the Ho, test statistic is

$$\chi^2 = \frac{ns^2}{\sigma^2} \text{ follows } \chi^2\text{-distribution with } (n-1) \text{dof}$$

where, s^2= variance sample

$$\frac{1}{n} \cdot \Sigma(x-\bar{x})$$

n = sample size.

s = standard deviation.

σ = Expected S.D.

σ^2 = Expected variance.

1) Weights in kg of 10 students are given below.

38, 40, 45, 53, 47, 43, 55, 48, 52, 49.

Can we say that variance of distribution of weights of all students from which the above sample of 10 students was drawn is equal to 20.

Soln: we set up the null hypothesis as $H_0 = \sigma^2 = 20$.

Calculation of sample variance.

x	38	40	45	53	47	43	55	48	52	49.
$x-\bar{x}$	-9	-7	-2	6	0	-4	8	1	5	2
$(x-\bar{x})^2$	81	49	4	36	0	16	64	1	25	4

$$\bar{x} = \frac{\Sigma x}{x} = \frac{470}{10} = 47.$$

Under the test statistic is $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{\Sigma(x-\bar{x})^2}{\sigma^2} = \frac{280}{20} = 14.$

which follows χ^2 distribution with dof $(10-1) = 9$.

Tabulated of χ^2 at 9 dof is 16.919. Since, calculated value of χ^2 is less than the tabulated value of for 9 dof at 5% level of significance. ~~It~~ is not significant; Hence, Ho may be accepted.

Subject : Date :

Title of the test case :

Case study No. : Page No. :

2) A Ramdom sample of size 20 form a population gives the sample standard deviation of 6. Test the hypothesis that the population S.D is 9.

Sol: We set up the null the hypothesis as

H_0 = The population standard deviation.

We are given $n=20$ and $s=6$.

under H_0: the test statistic is

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{20\times36}{81} = 8.89$$

and it follows χ^2-distribution $(20-1) = 19$ dof.

Tabulated value of χ^2 for 19 dof = 30.144.

Since, calculated value is less than the tabulated value, it is not significant. Hence, null hypothesis that the population standard deviation is 9 may be accepted at 5% level of significance.

chy-square test of goodness of fit :-

We are given a set of observed frequencies obtained under some experiment and we want to test if the experimental results support a particular hypothesis (or) theory.

Karl Pearson in 1900, developed a test for testing the significance of the decrepency between experimental values and the theoretical values obtained under some theory of hypothesis. This text known as f^2- Test of goodness of fit.

we set up the null hypothesis as there is no significant difference between the observed. (Experimental) and the theoretical (hypothetical) values.

Steps for consumption of f^2 f^2 and drawing the conclusions:

*1 Compute the expected frequencies $E_1, E_2, --- E_n$ Corresponding to the observed frequencies $O_1, O_2, --- O_n$. Under some theory of hypthesis.

*2 Compute the deviations $(O-E)$ for each frequency and then square them to obtain $(O-E)^2$.

*3 Divide the square of the deviations $(O-E)^2$ by the corresponding expected frequency to obtain $\frac{(O-E)^2}{E}$

*4 Add the values obtained in step(3) to compute

$$f^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

*5 Look at the tabulated values of f^2 fis $(n-1)$ dof at certain level of significance, usually 5% (or) 1% from the table of Significant values of f^2.

*6 If calculated value of f^2 is the less than the tabulated value, then it is said to be non-singnificant at the required level of significance and we may conclude that there is a good correspondance between theory & experiment.

*7 If calculated value of f^2 is greater than the tabulated value, it is said to be significant and we may conclude that the experiment does not support the theory.

1) The number of automobile accidents per week in a certain community were as follows.

12, 8, 20, 2, 14, 10, 15, 6, 9, 4.

Are these frequencies in agreement with the belief that accident conditions were the same during this 10-week period.

Sol: We set up the null hypothesis as the given frequencies are consistent with the belief that the accident conditions were same during the 10-week period.

Since, the total number of accidents over the 10-weeks are giv

$\therefore 12+8+20+2+14+10+15+6+9+4 = 100.$

Under the null hypothesis, these accidents should be uniformly distributed over the 10-week period and hence the expected number of accidents for each of the 10 weeks are $\frac{100}{10} = 10$.

Week	observed No. of accidents (O)	Expected No. of accidents (E)	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
1	12	10	2	4.	0.4
2	8	10	-2	4	0.4
3	20	10	10	100	10
4	2	10	-8	64	6.4.

5	14	10	4	16	1.6
6	10	10	0	0	0
7	15	10	5	25	2.5
8	6	10	-4	16	1.6
9	9	10	-1	1	0.1
10	4	10	-6	36	3.6
					26.6.

$$\therefore \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

$$= 26.6.$$

dof = 10-1=9, Tabulated $\chi^2_{0.05}$ for 9 dof = 16.919.

Since, calculated value $\chi^2 = 26.6$ is greater than the tabulated value 16.919, it is significant and null hypothesis is rejected at 5% level of significance.

2) In a mendelian experiment on breeding for types of plants are expected to occur in the proportion of 9:3:3:1. The observed frequencies are 891 round and yellow, 316 wrinkled and yellow, 290 round and green, and 119 wrinkled and green. Find the chy-square value and examine the correspondance between the theory and the experiment.

Sol:- We set up the null hypothesis as,

H₀: it is assumed that the theoretical values correspond to the experiment values.

Total no. of observed plans: 891+316+290+119=1616.

Expected Frequencies :-

Round & yellow $\Rightarrow \frac{9}{16} \times 1616 = 909$.

wrinkled & yellow $\Rightarrow \frac{3}{16} \times 1616 = 303$.

Round & Green $\Rightarrow \frac{3}{16} \times 1616 = 303$.

wrinkled & green $\Rightarrow \frac{1}{16} \times 1616 = 101$.

procedure is same $\chi^2 = 4.6799$.

dof $= 4-1 = 3$, Tabulated $\chi^2_{0.05}$ for 3dof $= 7.80$.

Since, calculated value of $\chi^2 = 4.6799$ is less than the tabulated value 7.80, it is not significant and null hypothesis is accepted at 5% level of significance.

Chy-square test for independence of attributes :-

Suppose that the given population consisting of N items is divided into 'r' mutually disjoint (Exclusive) & exhaustive classes $A_1, A_2, \ldots A_r$, with respect to the attribute 'A'.

Similarly, let us suppose that the same population is divided into 's' mutually disjoint & exhaustive classes $B_1, B_2, \ldots B_s$, with respect to the another attribute 'B'.

We set up null hypothesis as the two attributes A and B are independent.

if $(A_i B_j)_0$ denote the expected frequency of (A_i, B_j) then.

$$(A_iB_j) = \frac{(A_i)(B_j)}{N}$$

$$i = 1,2,----r$$
$$j = 1,2,----s.$$

i.e., the expected frequency for any cell frequency can be obtained on multiplying the row totals and column totals in which the frequency occurs and dividing the product by the total frequency 'N'.

Applying χ^2= test of goodyness of fit, the statistic is

$$\chi^2 = \sum_i \sum_j \left[\frac{(A_iB_j) - (A_iB_j)_0}{(A_iB_j)_0} \right]^2$$ follows χ^2-distribution with $(r-1) \times (s-1)$ dof

r = rows value.

s = columns value.

1) A certain drug was administrated to 456 males out of a total 720 in a certain locality to test its efficiency against typhoid. The incidence of typhoid is shown below. Find out the effectiveness of the drug against the disease.

	A1	A2	
	Infection	No infection	Total
B1 administering the drug	144 (A1B1)	312 (A2, B1)	456
B2 without administering the drug	192 (A1 B2)	72 (A2, B2)	264
Total	336	384	720

Subject : Date :

Title of the test case :

Case study No. : Page No. :

Sol:- We set up the null hypothesis as the two attributes incidence of typhoid and the 'administration of the drug' are independent. In other words, the drug is not effective against the disease.

Under Ho, the expected frequencies are,

$$E(144) = \frac{336 \times 456}{720} = 212.8$$

$$E(192) = \frac{336 \times 264}{720} = 123.2$$

$$E(312) = \frac{384 \times 456}{720} = 243.2$$

$$E(72) = \frac{384 \times 264}{720} = 140.8$$

Computation of χ^2:-

observed Frequency 'O'	Expected frequency 'E'	O-E	$(O-E)^2$
144	212.8	-68.8	4733.44
192	123.2	68.8	4733.44
312	243.2	68.8	4733.44
72	140.8	-68.8	4733.44

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 4733.44\left[\frac{1}{212.8} + \frac{1}{123.2} + \frac{1}{243.2} + \frac{1}{140.8}\right]$$

$$= 4733.44[0.0047 + 0.0081 + 0.0041 + 0.0071]$$

$$= 4733.44 \times 0.0240$$

$$= 113.60256.$$

follows χ^2 - dof = $(r-1)(s-1) = (2-1)(2-1) = 1$

Tabulated value of $\chi^2_{0.05} = 3.841$.

Since, calculated value of χ^2 is very much greater than tabulated value, it is highly significant. Hence, the null hypothesis is rejected at 5% level of significance & we conclude that the drug is certainly effective in containing typhoid.

8) Data on the hair colour and the eye colour are given in the table. calculate the χ^2 value. Determine the association between the hair colour and the eye colour.

		Fair	Brown	Black	Total.
Eye colour	Blue	15	20	5	40
	Grey	20	20	10	50
	Brown	25	20	15	60
	Total	60	60	30	150.

Sol:- (i) We set up null hypothesis as the 2 attributes association between hair colour & eye colour on same.

Under Ho the expected frequency are

i) $E(15) = \frac{40\times60}{150} = 160.$

ii) $E(20) = \frac{50\times60}{150} = 20.$

iii) $E(25) = \frac{60\times60}{150} = 24$

iv) $E(20) = \frac{40\times60}{150} = 16$

(v) $E(20) = \frac{50\times60}{150} = 20$

(vi) $E(20) = \frac{60\times60}{150} = 24$

(vii) $E(5) = \frac{40\times30}{150} = 8$

(viii) $E(10) = \frac{50\times30}{150} = 10$

(ix) $E(15) = \frac{60 \times 30}{150} = 12.$

Compute χ^2 :-

observed Frequency (O)	Expected frequency (E)	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
15	16	-1	1	0.0625
20	20	0	0	0
25	24	1	1	0.0417
20	16	4	16	1
20	20	0	0	0
20	24	-4	16	0.6667
5	8	3	9	1.125
10	10	0	0	0
15	12	3	9	0.7500
				3.6459

≃3.65.

Test statistic $\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$

$= 3.65$

we can assume that
level of significance is 5% i.e., 0.05.
degree of freedom (dof) = $r-1 \times s-1$
$= 3-1 \times 3-1$
$= 2 \times 2$
$= 4$

Tabulated values at 4 degree of freedom with 5% level of significance is 9.488.

Conclusion :-

Here, table value is high when compared to calculated value (3.65) i.e., 9.488 high than 3.65, So, the project is rejected at 4 degree of freedom: with 5% level of significance.

* According to rates correction ⇒ (2x2)

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Sign test for paired data :-

The sign test is the oldest of all non-parametric procedures and it was introduced by "Arbuthnott" (1710).

The sign test gets its name from the fact that it uses plus and minus signs rather than quantitative measurements as its data.

It is particularly useful where quantitative measurement is impossible (or) infeasible.

Applying z-test statistic to test the null hypothesis is $H_0: p = 1/2$ that

$$z = \frac{|r - n/2|}{\sqrt{n/4}}$$

where, r = no. of positive signs.

n = Total no. of signs [Except '0'].

Single data (or) ordinary sign test :-

1) The following data in turns are the amounts of sulphur oxides emitted by a large Industrial plant, in 40 days sample values :

17, 15, 20, 29, 19, 18, 22, 25, 27, 9, 24, 20, 17, 6, 24, 14, 15, 23, 24, 26, 19, 23, 28, 19, 16, 22, 24, 17, 20, 13, 19, 10, 23, 18, 31, 13, 20, 17, 24, 14.

Test the null hypothesis $\mu_0 = 21.5$.

Sol: Let $x_1, x_2, ---- x_n$ be the values of the sample size 'n' we want to test $H_0 = \mu = \mu_0$.

compare each of the 'n' values of the sample with μ_0.
If the difference is positive write '+' sign, if it is negative write '−' sign. If the difference is zero ignore all such values and read just the sample size 'n'.

sign of differences when compare with $\mu = 21.5$.

−, −, −, +, −, −, +, +, +, −, +, −, −, −, +,
−, −, +, +, +, −, +, +, −, −, +, +, −, −, −
−, −, +, −, +, −, −, −, +, −

r = no. of positive signs = 16

we want to test $H_0 : \mu_0 = 21.5$

under H_0, the test statistic is $\left(z_1 = \frac{|r - n/2|}{\sqrt{n/4}} \right) = \frac{|16-20|}{\sqrt{16}}$

critical value of z_1 at 5% is 1.96.

we accept the H_0.

Paired Data:

1) The nutritionist and medical doctors are always believed that vitamin c is highly effective in reducing the incidents of cold. To test this belief, a random sample of

13 persons is selected and they are given large daily doses of vitamin c under medical supervision over a period of 1 year. The number of persons who catch cold during the year is recorded and a comparision is made with the number of cold contacted by each such person daily the pervious year. This comparision is recorded as follows, the along with the sign of the change.

observations	1	2	3	4	5	6	7	8	9	10	11	12	13
with vitamin c	2	1	0	1	3	2	3	5	1	4	4	3	4
without vitamin c	7	5	2	3	8	2	4	4	3	7	6	2	10

Using the sign test at $\alpha = 0.05$ level of significance test whether vitamin c is effective in reducing the cold.

Soln: Let us table the null hypothesis that large is no difference in the number of cold contacted with (or) without vitamin 'c'.

Without vitamin c	7	5	2	3	8	2	4	4	3	7	6	2	10
With vitamin c	2	1	0	1	3	2	3	5	1	4	4	3	4
Sign	−	−	−	−	−	0	−	+	−	−	−	+	−

[To compare with '2' one first one is the bigger value at the time taken the sign is '−']

r = no. of positive signs = 2.

n = 12 [total signs Except '0']

under H_0 test statistic is $Z = \dfrac{\left|r - \frac{n}{2}\right|}{\sqrt{n/4}}$

Subject :

Date :

Title of the test case :

Case study No. :

Page No. :

$$= \frac{\left|2 - \frac{12}{2}\right|}{\sqrt{12/4}} = \frac{|2-6|}{\sqrt{3}} = \frac{4}{\sqrt{3}} = \frac{4}{1.73} = 2.31.$$

Since, calculated value $Z = 2.31$ is greater than the critical value $Z = 1.96$ at 5% level of significance they H_0 is rejected.

Code: 21E00105

MBA I Semester Regular Examinations May 2022

STATISTICS FOR MANAGERS

(Common to all)

(For students admitted in 2021 only)

Time: 3 hours Max. Marks: 60

All questions carry equal marks

Use of statistical tables is permitted.

*****

SECTION – A

(Answer the following: 05 X 10 = 50 Marks)

1 (a) Explain in brief the significance of statistics to business 5M

(b) The mean of marks in statistics of 100 students of a class was 72. The mean of marks of boys was 75, while their number was 70. Find out the mean marks of girls in the class. 5M

OR

2 An analysis of production rejects resulted in the following figures: 10M

No of rejects per operator	No of operators
21-25	5
26-30	15
31-35	28
36-40	42
41-45	15
46-50	12
51-55	3

Calculate mean and standard deviation.

3 The following data related to advertisement expenditure(in lakh of rupees) and their corresponding sales(in crore of rupees): 10M

Advertisement expenditure :	10	12	15	23	20
Sales :	14	27	23	25	21

Estimate:

(i) The sales corresponding to advertising expenditure of Rs.30 lakh.

(ii) The advertisement expenditure for a sales target of Rs.35 crore.

OR

4 (a) What is regression? What are its uses? 5M

(b) Calculate Karl Pearson's coefficient of correlation from the following data and interpret its value: 5M

Roll No.of students:	1	2	3	4	5
Marks in cost management:	48	35	17	23	47
Marks in statistics:	45	20	40	25	45

Contd. in page 2

5 A market survey conducted in 4 cities pertained to preference for brand A soap. The responses are shown below: 10M

	Delhi	Kolkata	Chennai	Mumbai
Yes	45	55	60	50
No	35	45	35	45
No opinion	5	5	5	5

(i) What is the probability that a consumer selected at random preferred brand A?

(ii) What is the probability that a consumer preferred brand A and was from Chennai?

(iii) What is the probability that a consumer preferred brand A, given that he/she was from Chennai?

(iv) Given that a consumer preferred brand A, what is the probability that he/she was from Mumbai?

OR

6 (a) What is conditional probability? Explain 5M

(b) If the probability of a defective bolt is 0.1, find: (i) The mean. (ii) The standard deviation of defective bolts in a total of 900. 5M

7 The sales data of an item in six shops before and after a special promotional campaign are as under: 10M

Shops:	A	B	C	D	E	F
Before campaign:	53	28	31	48	50	42
After campaign:	58	29	30	55	56	45

Can the campaign be judged to be a success? Test 5% level of significance

OR

8 (a) What is analysis of variance? What are the assumptions in analysis of variance? 5M

(b) In a big city 325 men out of 600 men were found be smokers. Does this information support the conclusion that the majority of men in this city are smokers? 5M

9 In an experiment on immunization of cattle from tuberculosis, the following results were found: 10M

	Affected	Not affected
Inoculated	12	26
Not inoculated	16	6

Calculate chi-square and discuss the effect of vaccine in controlling susceptibility to tuberculos is 5% level value of chi-square for one degree of freedom = 3.84)

OR

10 In a survey of 200 boys of which 75 were intelligent, 40 had educated fathers, while 85 of the unintelligent boys had uneducated fathers. Do these figures support the hypothesis that educated fathers have intelligent boys? (value of chi-square for 1 d.f is 3.84) 10M

SECTION – B

(Compulsory question, 01 X 10 = 10 Marks)

11 **Case Study/Problem:**

A certain company had 4 salesmen P, Q, R and S, each of whom was sent for a week into three types of areas A, B and C. The sales in kg per week are shown below: 10M

	Salesman			
District	P	Q	R	S
A	30	70	30	30
B	80	50	40	70
C	100	60	80	80

Carry out the analysis of variance and interpret the results.

*****

Code: 21E00105

MBA I Semester Supplementary Examinations October 2022

STATISTICS FOR MANAGERS

(Common to all)

(For students admitted in 2021 only)

Time: 3 hours Max. Marks: 60

All questions carry equal marks

*****

SECTION – A

(Answer the following: 05 X 10 = 50 Marks)

1 (a) List out any three differences between mean, median and mode. 5M

(b) Calculate the average bonus paid per member from the following data: 5M

Bonus(in Rs)	50	60	70	80	90	100	110
No. of workers	1	3	5	7	6	2	1

OR

2 What is meant by dispersion? In your opinion which is the best of finding out dispersion and why? 10M

3 (a) What is rank correlation? What are its merits? 5M

(b) Distinguish between correlation and regression. 5M

OR

4 The following data relate to the age of 10 employees and the number of days which they reported sick in a month. 10M

Age	20	30	32	35	40	46	52	55	58	62
Sick days	11	12	10	13	14	16	15	17	18	19

Calculate Karl Pearson's coefficient of correlation and interpret its value.

5 (a) Explain the Bayes' Theorem. 5M

(b) A bag contains 5 white and 4 black balls. Two balls are drawn at random one after the other without replacement. Find out the probability that both balls drawn are black. 5M

OR

6 The Human Resource department of a company has records which show the following analysis of its 200 engineers: 10M

Age	Bachelor's degree	Master's degree	Total
Under 30	90	10	100
30-40	20	30	50
Over 40	40	10	50

If one engineer is selected at random from the company, find:

(i) The probability he has only a bachelor's degree.

(ii) The probability he has a master's degree, given that he is over 40.

(iii) The probability he is under 30, given that he has only a bachelor's degree.

Contd. in page 2

7 Two types of drugs were used on 5 and 7 patients for reducing their weight. 10M
Drug A was imported and drug B was indigenous. The decrease in the weight after using the drugs for six months was as follows:

Drug A	10	12	13	11	14		
Drug B	8	9	12	14	15	10	9

Is there a significant difference in the efficacy of the two drugs? If not, which drug should you buy?(v = 10, t 0.05 = 2.223)

OR

8 Two random samples were drawn from the two normal populations and their values are: 10M

A	66	67	75	76	82	84	88	90	92		
B	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at the 5% level of significance (F=3.36) at 5% level for v_1 = 10 and v_2 = 8.

9 Of the 1,000 workers in a factory exposed to an epidemic, 700 in all were attacked, 400 had been inoculated and of these, 200 were attacked. On the basis of this information, can it be said that inoculation and attack are independent? 10M

	Inoculated	Not inoculated	Total
Attacked	200	500	700
Not attacked	200	100	300
Total	400	600	1,000

On the basis of this information, can it be said that inoculation and attack are independent? Carry out the chi-square test as per testing procedure at 5% level.

OR

10 Use the sign test to see if there is any difference between the number of days until collection of an account receivable before and after a new collection policy. Use the 0.05 significance level. 10M

Before: 30 28 34 35 40 42 33 38 34 45 28 27 25 41 36
After: 32 29 33 32 37 43 40 41 37 44 27 33 30 38 36

SECTION – B

(Compulsory question, 01 X 10 = 10 Marks)

11 **Case Study/Problem:** 10M

A test was given to 5 students chosen at random from the MBA class of each of the three universities in Andhra Pradesh. Their scores were found to be as follows:

University	Scores				
A	90	70	60	50	80
B	70	40	50	40	50
C	60	50	60	70	60

Carry out Analysis of Variance and show if there is any significant difference between the scores of students in the three universities (Given F5% = 3.44).

*****

Code: 17E00105

MBA I Semester Supplementary Examinations October 2020

STATISTICS FOR MANAGERS

(For students admitted in 2017, 2018 & 2019 only)

Time: 3 hours Max. Marks: 60

All questions carry equal marks

*****

SECTION – A

(Answer the following: 05 X 10 = 50 Marks)

1 (a) What are the uses of statistics in business decision making?

(b) Discuss the merits and demerits of range, quartile deviation and mean deviation.

OR

2 (a) What is meaning of standard deviation? Explain why standard deviation is the most preferred and widely used tool of measure of dispersion.

(b) An HR manager of a company finds that teenagers frequently change jobs. The dissatisfaction with their present jobs is a major factor in the decision they make. Thus, she selects a sample of interviews of 15 teenagers from the past six months. She records the number of months the teenagers spent on their previous jobs:

12 5 1 6 20 24 16 7 11 8 23 19 25 14 4

(i) Calculate the range of months that the teenagers spent on their jobs.

(ii) Calculate the median months that each spent at their previous job.

(iii) Calculate the interquartile range for the months each teenager spent at his or her previous job.

3 (a) What is ‘correlation’? Explain positive and negative correlations.

(b) State the properties of regression coefficients.

OR

4 (a) Explain the concept of regression sum of squares (SSR) and error of squares (SSE) in a regression model.

(b) Consider the following set of data:

x	48	27	34	24	49	29	39	38	46	32
y	47	23	31	20	50	48	47	47	42	47

Calculate the correlation coefficient of these two variables.

5 (a) Explain the concept of normal distribution. Analyse why it is a widely used probability distribution.

(b) Define the mean, standard deviation and variance of an exponential distribution.

OR

6 (a) What is a Poisson distribution? State the main assumptions of a Poisson distribution.

(b) A consumer electronics company has 24 showrooms located across India. Out of these 24 showrooms, 12 are located in Gujarat. If five showrooms are selected at random from the entire list, what is the probability that one or more randomly selected showrooms are located in Gujarat?

Contd. in page 2

7 Discuss the procedure of testing hypothesis for difference between two sample means using t-test.

OR

8 Explain the different types of ANOVA. What are the steps involved in carrying out ANOVA?

9 (a) What is the difference between parametric tests and non-parametric tests?

(b) Discuss the procedure of conducting Chi-square test for independence of attributes.

OR

10 (a) Explain the assumptions and significance of sign test.

(b) The table below gives the scores obtained from a random sample of 8 customers before and after the demonstration of a product. Is there any evidence of difference in scores before and after demonstration?

Scores before product demonstration	Scores after product demonstration
30	28
32	40
31	44
34	30
30	41
32	42
34	43
31	29

SECTION – B

(Compulsory question, 01 X 10 = 10 Marks)

11 **Case Study/Problem:**

A company is concerned about the high rates of absenteeism among its employees. It organised a training programme to boost the morale of its employees. The following table gives the number of days that sixteen randomly selected employees have received training, and the number of days they have availed leave:

Employee	Training days	Leave
1	12	20
2	14	18
3	16	16
4	13	22
5	11	18
6	10	19
7	15	14
8	17	12
9	18	10
10	19	9
11	17	11
12	15	16
13	13	19
14	15	17
15	17	15
16	12	21

Questions:

(i) Develop a regression model to predict leaves based on training days.

(ii) Calculate the coefficient of determination and interpret it.

(iii) Calculate the standard error of the estimate.

(iv) Predict the leaves when the training days are 25.

*****

Code: 17E00105

MBA & MBA (Finance) I Semester Regular & Supplementary Examinations January 2020

STATISTICS FOR MANAGERS

(For students admitted in 2017, 2018 & 2019 only)

Time: 3 hours Max. Marks: 60

All questions carry equal marks

*****

SECTION – A

(Answer the following: 05 X 10 = 50 Marks)

1 (a) Explain the nature and significance of statistics to business.

(b) What are the differences among the mean, median and mode?

OR

2 (a) State the business applications of measures of central tendency.

(b) The numbers of work stoppages in a country over the last 10 years are 22, 20, 21, 15, 5, 11, 19, 19, 15, and 11.

(i) What is the median number of stoppages?

(ii) How many observations are below the median? Above it?

(iii) What is the modal number of work stoppages?

3 (a) What is the significance of correlation?

(b) The following sample of observations was randomly selected:

x	5	3	6	3	4	4	6	8
y	13	15	7	12	13	11	9	5

Determine the correlation coefficient and interpret the relationship between *x* and *y*.

OR

4 (a) Explain different types of correlations.

(b) State the assumption of regression analysis.

5 (a) What do you understand by the following in the theory of probability:

(i) Mutually exclusive events.

(ii) Collectively exhaustive events.

(b) A store receives 3 red, 6 white, and 7 blue shirts. Two shirts are drawn at random. Determine the probability that:

(i) Both the shirts are white.

(ii) One shirt is red and the other shirt is white.

(iii) Both the shirts are blue.

OR

6 (a) Explain the concept of conditional probability.

(b) A company employed 150 employees of whom 40 are mechanical engineers and 110 are diploma holders in management. Thirty per cent of the management diploma holders are mechanical engineers. If an employee is selected at random, what is the probability that the employee is a management diploma holder and a mechanical engineer?

7 (a) Explain the properties and importance of the *F* distribution.

(b) Identify the limitations of two-way ANOVA?

OR

Contd. in page 2

8 (a) What is the importance of hypothesis testing in managerial decision making?

(b) Use the following data to perform one-way ANOVA.

Factor 1	Factor 2	Factor 3	Factor 4
13	17	22	18
12	15	26	17
13	18	27	16
14	16	28	15
15	17	29	16
13	18	30	17

Use α = 0.05 to test the hypotheses for the difference in means.

9 (a) Explain the conceptual framework of χ^2 test with respect to expected and observed frequencies.

(b) Discuss the concept of contingency table.

OR

10 (a) Explain the major advantages of non-parametric tests over parametric tests.

(b) A sample of 300 employees who own a car among various IT companies were questioned on whether they own a luxury car. The responses obtained were tabulated as given below:

Company	Infosys	Wipro	TCS	CTS	Tech Mahindra	Accenture	IBM	Row Total
Employees owning a luxury car	10	4	7	8	3	9	9	50
Employees not owning a luxury car	30	31	38	22	27	56	46	250
Column total	40	35	45	30	30	65	55	300

(i) Find the degree of freedom for conducting the chi-square test.
(ii) Develop a table of observed frequency and expected frequency.
(iii) Find the chi-square value for the given data.

SECTION – B

(Compulsory question, 01 X 10 = 10 Marks)

11 **Case Study:**

Problem:

A large supermarket has adopted a new strategy to increase its sales. It has adopted a few consumer friendly policies and is using video clips of 15 minutes to propagate the new policies. The following table provides data about the number of video clips shown in a randomly selected day and the sales turnover of the supermarket in the corresponding day.

Days	No. of video clips shown	Sales (in thousand rupees)
1	25	150
2	25	210
3	25	140
4	35	180
5	35	230
6	35	270
7	40	310
8	40	330
9	40	300
10	50	270
11	50	310
12	50	340

(i) Develop a regression model to predict sales from the number of video clips shown.
(ii) Calculate the coefficient of determination and interpret it.
(iii) Calculate the standard error of the estimate.

*****