JEPPIAAR
ENGINEERING COLLEGE

DEPARTMENT OF MANAGEMENT STUDIES

II YEAR / III SEMESTER

BA4101: STATISTICS FOR MANAGEMENT

STUDY MATERIAL

Faculty In charge

Dr. P. SIVAGAMI

Anna University Chennai

Regulation 2021

TABLE OF CONTENTS

JEPPIAAR
ENGINEERING COLLEGE

Jeppiaar Nagar, OMR Salai, Semmencherry ,Chennai -600119

VISION

To build Jeppiaar Engineering College as an institution of academic excellence in technology and management education, leading to become a world class university.

MISSION

- To excel in teaching and learning, research and innovation by promoting the principles of scientific analysis and creative thinking.
- To participate in the production, development and dissemination of knowledge and interact with national and international communities.
- To equip students with values, ethics and life skills needed to enrich their lives and enable them to contribute for the progress of society.
- To prepare students for higher studies and lifelong learning, enrich them with the practical skills necessary to excel as future professionals and entrepreneurs for the benefit of Nation's economy.

DEPARTMENT OF MANAGEMENT STUDIES

VISION

To be a prominent management institution developing industry ready managers, entrepreneurs and socially responsible leaders by imparting extensive expertise and competencies.

MISSION

- To provide management education to all groups in the community.
- To practice management through scholarly research and education.
- To advance in the best practices of management which enable the students to meet the global industry demand.
- To promote higher studies, lifelong learning, entrepreneurial skills and develop socially responsible professionals for empowering nation's economy.

PROGRAMME EDUCATIONAL OBJECTIVES (PEOs):

MBA programme curriculum is designed to prepare the post graduate students

- To have a thorough understanding of the core aspects of the business.
- To provide the learners with the management tools to identify, analyze and create business opportunities as well as solve business problems.
- To prepare them to have a holistic approach towards management functions.
- To inspire and make them practice ethical standards in business.

PROGRAMME OUTCOMES (POs)

On successful completion of the programme,

1. Ability to apply the business acumen gained in practice.
2. Ability to understand and solve managerial issues.
3. Ability to communicate and negotiate effectively, to achieve organizational and individual goals.
4. Ability to understand one's own ability to set achievable targets and complete them.
5. Ability to fulfill social outreach
6. Ability to take up challenging assignments

COURSE OBJECTIVE:

 To learn the applications of statistics in business decision making.

COURSE OUTCOMES:

 C101.1: Facilitate objective solutions in business decision making.
 C101.2: Understand and solve business problems.
 C101.3: Apply statistical techniques to data sets, and correctly interpret the results.
 C101.4: Develop skill-set that is in demand in both the research and business environments.
 C101.5: Enable the students to apply the statistical techniques in a work setting.

CO-PO Matrix

CO	PO1	PO2	PO3	PO4	PO5	PO6
CO1	3	3	3	0	0	2
CO2	3	3	3	0	0	2
CO3	3	3	3	0	0	2
CO4	3	3	3	0	0	2
CO5	3	3	3	0	3	2
Average	**3**	**3**	**3**	**0**	**3**	**2**

BA4101 - STATISTICS FOR MANAGEMENT

Syllabus

COURSE OBJECTIVE:

To learn the applications of statistics in business decision making.

UNIT I INTRODUCTION

Basic definitions and rules for probability, conditional probability independence of events, Baye‘s theorem, and random variables, Probability distributions: Binomial, Poisson, Uniform and Normal distributions.

UNIT II SAMPLING DISTRIBUTION AND ESTIMATION

Introduction to sampling distributions, sampling distribution of mean and proportion, application of central limit theorem, sampling techniques. Estimation: Point and Interval estimates for population parameters of large sample and small samples, determining the sample size.

UNIT III TESTING OF HYPOTHESIS – PARAMETIRC TESTS

Hypothesis testing: one sample and two sample tests for means and proportions of large samples (ztest), one sample and two sample tests for means of small samples (t-test), F-test for two sample standard deviations. ANOVA one and two way

UNIT IV NON-PARAMETRIC TESTS

Chi-square test for single sample standard deviation. Chi-square tests for independence of attributes and goodness of fit. Sign test for paired data. Rank sum test. Kolmogorov-Smirnov – test for goodness of fit comparing two populations. Mann – Whitney U test and Kruskal Wallis test. One sample run test.

UNIT V CORRELATION AND REGRESSION

Correlation – Coefficient of Determination – Rank Correlation – Regression – Estimation of Regression line – Method of Least Squares – Standard Error of estimate.

REFERENCES:

1. Richard I. Levin, David S. Rubin, Masood H.Siddiqui, Sanjay Rastogi, Statistics for Management, Pearson Education, 8th Edition, 2017.

2. Prem. S. Mann, Introductory Statistics, Wiley Publications, 9th Edition, 2015.

3. T N Srivastava and Shailaja Rego, Statistics for Management, Tata McGraw Hill, 3rd Edition 2017.

4. Ken Black, Applied Business Statistics, 7th Edition, Wiley India Edition, 2012.

5. David R. Anderson, Dennis J. Sweeney, Thomas A.Williams, Jeffrey D.Camm, James J.Cochran, Statistics for business and economics, 13th edition, Thomson (South – Western) Asia, Singapore, 2016.

6. N. D. Vohra, Business Statistics, Tata McGraw Hill, 2017.

UNIT I INTRODUCTION

Basic definitions and rules for probability, conditional probability independence of events, Baye's theorem, and random variables, Probability distributions: Binomial, Poisson, Uniform and Normal distributions.

UNIT-1 INTRODUCTION

Random experiment: An experiment whose all possible outcomes are known, but it is not possible to predict the outcome.

Probability:

Let A be a event and B be a its sample space then its probability on the occurrence on events is defined as $P(A)=\dfrac{No.\ of\ favourable\ Cases}{Total\ no.\ of\ exhaustic\ Cases}$.

Axioms of Probability:

(i) $0 \le P(E) \le 1$ (ii) $P(S)=1$ (iii) $P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$ if E$_i$'s are mutually exclusive events.

Example: (i) A fair coin is "tossed" (ii) A die is "rolled" are random experiments, since we cannot predict the outcome of the experiment in any trial.

Mutually exclusive:

Two events are said to be mutually exclusive if the occurrence of any one of them excludes the occurrence of other in a single experiment.

Example: Tossing of Coin.

Independent events:

Two (or) more events are independent if the occurrence of one does not affect the occurrence of the other.

Example: If coin is tossed twice; result of second throw is not affected by the result of first throw.

Addition Law of Probability:

If A and B are two events in a sample space "S" then

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

Conditional Probability:

The conditional probability of an event B assuming that the event A has happened, is defined as $P(B/A)=\dfrac{P(A\cap B)}{P(A)}, P(A)\neq 0$

Similarly, $P(A/B)=\dfrac{P(A\cap B)}{P(B)}, P(B)\neq 0$.

1. If A and B are independent events then a) $A\ and\ \overline{B}$ b) $\overline{A}\ and\ \overline{B}$ are also independent.

Solution:

Since A and B are independent,

$$P(A\cap B)=P(A)P(B)---1$$

a) $P(A\cap \overline{B})=P(A)-P(A\cap B)$

$$=P(A)-P(A)P(B) \quad [u\sin g(1)]$$

$$=P(A)\left[1-P(B)\right]$$

$$P(A\cap \overline{B})=P(A)P(\overline{B}) \quad \therefore A\,\&\,B\ are\ independent\ events$$

b) $P(\overline{A}\cap \overline{B})=P(\overline{A\cup B})$

$$=1-P(A\cup B)$$

$$=1-\left[P(A)+P(B)-P(A\cap B)\right]\ [By\ addition\ theorem]$$

$$=1-P(A)-P(B)+P(A\cap B)$$
$$=1-P(A)-P(B)+P(A)P(B) \quad [u\sin g-1]$$
$$=1-P(A)-P(B)[1-P(A)]$$
$$=[1-P(A)][1-P(B)]$$
$$P(\bar{A}\cap\bar{B})=P(\bar{A})P(\bar{B})$$
$\therefore \bar{A}\ \&\ \bar{B}$ *are indepentent events.*

2. A Problem in statistics is given to three students. A, B and C whose chances of solving it are $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{4}$ respectively. What is the probability that the problem will be solved?

Solution:

Let A, B, C Denote the events that the problem is solved by the students A, B, C respectively.

Then $P(A)=\frac{1}{2}$, $P(B)=\frac{1}{2}$, $P(C)=\frac{1}{4}$

$$P(\bar{A})=1-\frac{1}{2}=\frac{1}{2}$$
$$P(\bar{B})=1-\frac{1}{3}=\frac{2}{3}$$
$$P(\bar{C})=1-\frac{1}{4}=\frac{3}{4}$$

$\text{P(all the three students will not solve the problem)}=P(\bar{A})P(\bar{B})P(\bar{C})=\frac{1}{2}.\frac{2}{3}.\frac{3}{4}=\frac{1}{4}$

$\therefore \text{P(all the three students will solve the problem)}=P(A\cup B\cup C)$

$$=1-P(\bar{A})P(\bar{B})P(\bar{C})=1-\frac{1}{4}=\frac{3}{4}$$

3. Event A and B are such that $P(A+B)=\frac{3}{4}$, $P(AB)=\frac{1}{4}$ and $P(\bar{A})=\frac{2}{3}$ find $P(B)$.

Solution:

Given $P(A+B)=\frac{3}{4}, P(AB)=\frac{1}{4}, P(\bar{A})=\frac{2}{3}$

i.e. $P(A)=1-P(\bar{A})=1-\frac{2}{3}=\frac{1}{3}$

By addition theorem

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

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

$$P(B)=\frac{3}{4}-\frac{1}{3}+\frac{1}{4}=\frac{9-4+3}{12}=\frac{8}{12}=\frac{2}{3}$$

4. An integer is chosen at random from two hundred digits. What is the probability that the integer is divisible by 6 or 8?

Solution:

The sample space $=\{1,2,3\ldots\ldots199,200\}$

$$n(S)=100$$

Let the event A be an integer chosen that is divisible by 6,

i.e. $A=\{6,12,18\ldots\ldots\ldots198\}$

$$n(A)=\frac{198}{6}=33$$

$$\therefore n(A)=\frac{n(A)}{n(S)}=\frac{33}{200}$$

Let the event B be an integer chosen that is divisible by 8
i.e. $B=\{8,16,24.....200\}$

$$n(B)=\frac{200}{8}=25$$

$$\therefore P(B)=\frac{n(B)}{n(S)}=\frac{25}{200}$$

The L.C.M of 6 & 8 is 24.
Hence, a number that is divisible by both 6 & 8 is divisible by 24.

$$\therefore A\cap B=\{24,48,72,.....192\}$$

$$n(A\cap B)=\frac{192}{24}=8$$

$$\therefore P(A\cap B)=\frac{n(A\cap B)}{n(S)}=\frac{8}{200}$$

Hence by addition theorem on probability

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)=\frac{33}{200}+\frac{25}{200}-\frac{8}{200}=\frac{58-8}{200}=\frac{50}{200}=\frac{1}{4}$$

5. A **and** B **throw alternatively with a pair of dice.** A **wins if he throws 6 before** B **throws 7 and 8 wins if he throws 7 before a throws 6.If A begins, show that their respective chances of winning are in the ratio 30:61.**

Solution:

Let A_i denote the event of A's throwing 6 in the i^{th} thrown $i=1,2,3,\ldots$

'6' can be obtained with two dice in the following ways

$(1,5)(5,1)(2,4)(4,2)(3,3)$

i.e. 5 distinct ways

$$P(A_i)=\frac{5}{36},\ P(\overline{A_i})=1-P(A_i)=\frac{31}{36},\ i=1,2,\ldots$$

Let B_i denote the event of B's throwing 7 in the i^{th} thrown $i=1,2,3,\ldots$

'7' can be obtained with two dice in the following ways

$(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)$

i.e. 6 distinct ways

$$P(B_i)=\frac{6}{36},P(\overline{B_i})=1-P(B_i)=\frac{5}{6},\ i=1,2,.....$$

6. If A starts the game, he will win in the following mutually exclusive ways:

(i) A_1 happens (ii) $\overline{A_1}\cap\overline{B_2}\cap A_3$ happens

(iii) $\overline{A_1}\cap\overline{B_2}\cap\overline{A_3}\cap\overline{B_4}\cap A_5$ happens, and so on.

Hence by addition theorem of probability, the required probability of winning is given by $P(A)$,

$$\begin{aligned}P(A)&=P(i)+P(ii)+P(iii)+\ldots\\&=P(A_1)+P(\overline{A_1}\cap\overline{B_2}\cap A_3)+P(\overline{A_1}\cap\overline{B_2}\cap\overline{A_3}\cap\overline{B_4}\cap A_5)+\ldots\\&=P(A_1)+P(\overline{A_1})P(\overline{B_2})P(A_3)+P(\overline{A_1})P(\overline{B_2})P(\overline{A_3})P(\overline{B_4})P(A_5)+..\end{aligned}$$

[$\because$ the events are mutually independent]

$$=\frac{5}{36}+\left(\frac{31}{36}\times\frac{5}{6}\right)\times\frac{5}{36}+\left(\frac{31}{36}\times\frac{5}{6}\right)\times\left(\frac{31}{36}\times\frac{5}{6}\right)\times\frac{5}{36}+\ldots$$

$$\left[\because \text{ the series is an infinite geometric series } \frac{a}{1-r} \; where\; a=\frac{5}{36} \;\&\; r=\left(\frac{31}{36}\right)\left(\frac{5}{6}\right)\right]$$

$$=\frac{\frac{5}{36}}{\left(1-\frac{31}{36}\right)\times\frac{5}{6}}=\frac{\frac{5}{36}}{\frac{61}{216}}=\frac{30}{61}.$$

7. The probability that a contractor will get a plumbing contract is $2/3$ and the probability that he will not get an electric contract is $5/9$. If the probability of getting at least one contract is $4/5$, what is the probability that he will get both ?

Solution:

Let A be an event of getting a plumbing contract & B be an event of getting an electric contract.

$$P(A)=\frac{2}{3}, P(\overline{B})=\frac{5}{9}, P(A\cup B)=\frac{4}{5}$$

$$P(B)=1-P(\overline{B})=1-\frac{5}{9}=\frac{4}{9}$$

By addition theorem of probability

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

$$P(A\cap B)=\frac{2}{3}+\frac{4}{9}-\frac{4}{5}=\frac{30+20-36}{45}=\frac{50-36}{45}=\frac{14}{45}$$

i.e. probability of getting both the contract is $\frac{14}{15}$.

8. Let $P(A\cup B)=\frac{5}{6}, P(A\cap B)=\frac{1}{3}$ *and* $P(\overline{B})=\frac{1}{2}$. Are the events A and B independent Explain.

Solution:

$$P(B)=1-P(\overline{B})=\frac{1}{2}$$

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

$$=\frac{5}{6}+\frac{1}{3}-\frac{1}{2}=\frac{5+2-3}{6}=\frac{4}{6}=\frac{2}{3}$$

Since $P(A\cap B)=\frac{1}{3}=P(A)P(B)$ $\quad\because P(A)P(B)=\frac{2}{3}\frac{1}{2}=\frac{1}{3}$

Hence A & B are independent

Total probability of an event:

If $A_1, A_2, ..., A_n$ are mutually exclusive and exhaustive events and B is any event in S then $P(B)=P(B_1)P(B/A_1)+P(B_2)P(B/A_2)+...+P(B_n)P(B/A_n)$.

State Baye's theorem.

If $E_1, E_2, ..., E_n$ are mutually disjoint events with $P(E_i)\neq 0, (i=1,2,...,n)$ then for any arbitrary event a which is a subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A)>0$ We have

$$P(E_i/A)=\frac{P(E_i)P(A/E_i)}{\sum_{i=1}^{n}P(E_i)P(A/E_i)}, i=1,2,...,n.$$

9. If the probability that A solves a problem is $1/2$ and that for B is $3/4$ and if they

aim at solving a problem independently, what is the probability that the problem is solved?

Solution:

Probability of A solving a problem is $P(A)=\frac{1}{2}$ & that of B is $P(B)=\frac{3}{4}$.

$\because A \& B$ are independent $P(A\cap B)=P(A)P(B)=\frac{1}{2}.\frac{3}{4}=\frac{3}{8}$

Hence the probability that the problem is solved is

$P(A\cup B)=P(A)+P(B)-P(A\cap B)=\frac{1}{2}+\frac{3}{4}-\frac{3}{8}$ Using (1)

$$=\frac{4+6-3}{8}=\frac{10-3}{8}=\frac{7}{8}\ .$$

10. In a shooting test, the probability of hitting the target is $\frac{1}{2}$ for A, $\frac{2}{3}$ for B and $\frac{3}{4}$ for C. If all of them fine at the target, find the probability that (i) none of them hits the target and (ii) atleast one of them hits the target.

Solution:

Given $P(A)=\frac{1}{2}$, $P(B)=\frac{2}{3}$, $P(C)=\frac{3}{4}$

$$P(\overline{A})=\frac{1}{2}\ , P(\overline{B})=\frac{1}{3}, P(\overline{C})=\frac{1}{4}$$

(i) $P(\overline{A}\cap\overline{B}\cap\overline{C})=P(\overline{A})P(\overline{B})P(\overline{C})$ (by independence) $=\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}=\frac{1}{24}$

(ii) P (atleast one hits the target) =1−P(none hits the target) $=1-\frac{1}{24}=\frac{23}{24}$

11. A bolt is manufactured by 3 machines A,B and C. A turns out twice as many items as B, and machines B and C produce equal number of items. 2% of bolts produced by A and B are defective and 4% of bolts produced by C are defective. All bolts are put into 1 stock pile and 1 is chosen from this pile. What is the Probability that it is defective?

Solution:

Let A, B & C be the event in which the item has been produced by machine A, B & C respectively.

D be the event of the item being defective.

Given $P(A)=\frac{1}{2}, P(B)=P(C)=\frac{1}{4}$

$P(D/A)=P$ (an item is defective, given that A has produced it)

$$=\frac{2}{100}=P(D/B)$$

$$P(D/C)=\frac{4}{100}$$

By theorem of total probability,

$$P(D)=P(A)\times P(D/A)+P(B)\times P(D/B)+P(C)\times P(D/C)$$

$$=\frac{1}{2}\times\frac{2}{100}+\frac{1}{4}\times\frac{2}{100}+\frac{1}{4}\times\frac{4}{100}\quad=\frac{1}{100}+\frac{1}{200}+\frac{1}{100}=\frac{2}{100}+\frac{1}{200}$$

$$=\frac{4+1}{200}=\frac{5}{200}=\frac{1}{40}.$$

12. For a certain binary communication channel, the probability that a transmitted '0' is received as a '0' is 0.95 and the probability that a transmitted '1' is received as '1'is 0.90. If the probability that (i) a '1' is received and (ii) a '1' was transmitted given that a '1' was received.

Solution:

Let A be the event of transmitting '1'

$\overline{A}$ be the event of transmitting '0'
B be the event of received '1' &
$\overline{B}$ be the event of receiving '0'.

Given $P\left(\overline{A}\right)=0.4, P(B/A)=0.9$ & $P\left(\overline{B}/\overline{A}\right)=0.95$

$\therefore P(A)=1-P\left(\overline{A}\right)=1-0.4=0.6$ & $P\left(B/\overline{A}\right)=0.05$

By the theorem of total probability

$$P(B)=P(A)P(B/A)+P\left(\overline{A}\right)P\left(B/\overline{A}\right) = 0.6\times 0.9+0.4\times 0.05 = 0.56$$

By Baye's Theorem,

$$P(A/B)=\frac{P(A)\times P(B/A)}{P(B)}=\frac{0.6\times 0.9}{0.56}=\frac{27}{28}.$$

13. A given lot of IC chips contains 2% defective chips. Each chip is tested before delivery. The tester itself is not reliable. Probability of tester says the chip is good when it is really good is 0.95 and the probability of tester says chip is defective when it is actually defective is 0.94. If a tested device is indicated to be defective, what is the probability that it is actually defective?

Solution:

Let A be the event of chip that is actually defective & B be the event of chip that is actually good.

Let D be the event of tester says it is good.

Given $P(A)=0.02$, $P(B)=0.98$, $P(D/B)=0.95$

$$P\left(\bar{D}/A\right)=0.94,\ P\left(\bar{D}/B\right)=1-P(D/B)=1-0.95=0.05$$

By the theorem of total probability

$$P\left(\bar{D}\right)=P(A)P\left(\bar{D}/A\right)+P(B)P\left(\bar{D}/B\right)=0.02\times 0.94+0.98\times 0.05 = 0.0678$$

By Baye's rule,

$$P\left(A/\bar{D}\right)=\frac{P(A)P\left(\bar{D}/A\right)}{P\left(\bar{D}\right)}=\frac{0.94\times 0.02}{0.0678}$$

$$P\left(A/\bar{D}\right)=0.2772.$$

14. An urn contains 5 balls. Two balls are drawn and are found to be white. What is the probability of all the balls being white?

Solution:

A_1	A_2	A_3	A_4
W O	W O	W O	W
2 3	3 2	4 1	5

Where W denotes white ball & O denotes other colors.

Let it be A_1, A_2, A_3, A_4

$$\therefore P(A_1)=P(A_2)=P(A_3)=P(A_4)=\frac{1}{4}$$

Let D be the event of selecting 2 white balls

$$P(D/A_1)=\frac{2C_2}{5C_2}=\frac{1}{10}$$

$$P(D/A_2)=\frac{3C_2}{5C_2}=\frac{3}{10}$$

$$P(D/A_3)=\frac{4C_2}{5C_2}=\frac{3}{5}$$

$$P(D/A_4)=\frac{5C_2}{5C_2}=1$$

By the theorem of total probability

$$P(D)=P(A_1)P(D/A_1)=P(A_2)P(D/A_2)+P(A_3)P(D/A_3)+P(A_4)P(D/A_4)$$

$$=\frac{1}{4}\left(\frac{1}{10}+\frac{3}{10}+\frac{3}{5}+1\right)=\frac{1}{4}\left(\frac{1+3+6+10}{10}\right)=\frac{20}{40}=\frac{1}{2}$$

By Bayer's theorem,

$$P(A_4/D)=\frac{P(A_4)P(D/A_4)}{P(D)}=\frac{\frac{1}{4}\times 1}{\frac{1}{2}}=\frac{1}{2}.$$

15. The first bag contains 3 white balls, 2 red balls and 4 black balls. Second bag contains 2 white, 3 red and 5 black balls and third bag contains 3 white, 4 red and 2 black balls. One bag is chosen at random and from it 3 balls are drawn. Out of three balls two balls are white and one is red. What are the probabilities that they were taken from first bag, second bag, third bag.

Solution:

Urn I			*Urn II*			*Urn III*		
3	2	4	2	3	5	3	4	2
W	*R*	*B*	*W*	*R*	*B*	*W*	*R*	*B*

Where W, R, B denotes white, red & black balls

Let it be $A_1, A_2,\ A_3$

$$P(A_1)=P(A_2)=P(A_3)=\frac{1}{3}$$

Let D be the event of selecting three balls taken from the selected bag that 2 are white and 1 is red.

$$P(D/A_1)=\frac{3C_2\times 2C_1}{9C_3}=\frac{6}{84}$$

$$P(D/A_2)=\frac{2C_2\times 3C_1}{10C_3}=\frac{3}{120}$$

$$P(D/A_3)=\frac{3C_2\times 4C_1}{9C_3}=\frac{12}{84}$$

By the theorem of total probability,

$$P(D)=P(A_1)P(D/A_1)+P(A_2)P(D/A_2)+P(A_3)P(D/A_3)$$

$$=\frac{1}{3}\times\frac{6}{84}+\frac{1}{3}\times\frac{3}{120}+\frac{1}{3}\times\frac{12}{84}=0.0746.$$

By Baye's theorem,

$$P(A_1/B)=\frac{P(A_1)P(B/A_1)}{P(D)}=\frac{\frac{6}{84}\times\frac{1}{3}}{0.0746}=0.319$$

$$P(A_2/B)=\frac{P(A_2)P(B/A_2)}{P(D)}=\frac{\frac{1}{3}\times\frac{3}{120}}{0.0746}=0.0428$$

$$P(A_3/B)=\frac{P(A_3)P(B/A_3)}{P(D)}=\frac{\frac{1}{3}\times\frac{12}{84}}{0.0746}=0.6380$$

16. A factory produces a certain type of outputs by three types of machine. The respective daily Production figures are: Machine I: 3,000 Units; Machine II: 2,500 Units; Machine III: 4,500 Units.Past experience shows that I percent of the output produced by machine I is defective. The corresponding fractions of defectives for the other two machines are 1.2 percent and 2 percent respectively. An item is drawn at random from the day's production run and is

found to be defective. What is the probability that it comes from the output of (i) Machine I, (ii) Machine II, (iii) Machine III?

Solution:

Let A_1A_2 & A_3 denote the events that the output is produced by machines I, II & III respectively and let D denote the event that the output is defective.

Thus

$$P(A_1)=\frac{3000}{10,000}=0.30, P(A_2)=\frac{2500}{10,000}=0.25, P(A_3)=\frac{4500}{10,000}=0.45$$

$$P(D/A_1)=1\%=0.01, P(D/A_2)=1.2\%=0.012, P(D/A_3)=2\%=0.02$$

By the theorem of total probability,

$$P(D)=P(A_1)P(D/A_1)+P(A_2)P(D/A_2)+P(A_3)P(D/A_3)$$

$$=(0.30)(0.01)+(0.25)(0.012)+(0.45)(0.02)=0.015$$

By Baye's rule

(i) $P(A_1/D)=\dfrac{P(A_1)P(D/A_1)}{P(D)}=\dfrac{0.003}{0.015}=\dfrac{1}{5}$

(ii) $P(A_2/D)=\dfrac{P(A_2)P(D/A_2)}{P(D)}=\dfrac{0.003}{0.015}=\dfrac{1}{5}$

(iii) $P(A_3/D)=\dfrac{P(A_3)P(D/A_3)}{P(D)}=\dfrac{0.009}{0.015}=\dfrac{3}{5}$

17. There are two boxes B_1 and B_2 . B_1 contains two red balls and one green ball. B_2 contains one red ball and two green balls.

(i)A ball is drawn from one of the boxes randomly. It is found to be red. What is the Probability that it is from B_1?

(ii)Two balls are drawn randomly from one of the boxes without replacement. One is red and the other is green. What is the probability that they came from B_1?

(iii)A ball drawn from one of the boxes is green. What is the probability that it came from B_2?

(iv)A ball drawn from one of the boxes is white. What is the probability that it came from B_2?

Solution:

Let B_1 and B_2 be the events that the boxes B_1 and B_2 are selected respectively.

$$P(B_1)=\frac{1}{2}, P(B_2)=\frac{1}{2}$$

Let A be the event that a red ball is selected.

$$P(A/B_1)=\frac{2}{3}, P(A/B_2)=\frac{1}{3}$$

By Baye's theorem,

(i) P(ball is from B_1, given it is red)

$$=P(A/B_1)=\frac{P(B_1)P(A/B_1)}{P(B_1)P(A/B_1)+P(B_2)P(A/B_2)}=\frac{\frac{1}{2}\times\frac{2}{3}}{\left(\frac{1}{2}\times\frac{2}{3}\right)+\left(\frac{1}{2}\times\frac{1}{3}\right)}=\frac{2}{3}$$

(ii) Let C be the event that a red ball and a green ball are selected.

$$P(C/B_1)=\frac{2_{C_1}\times 1_{C_1}}{3_{C2}}=\frac{2}{3}, P(C/B_2)=\frac{1_{C_1}\times 2_{C_1}}{3_{C2}}=\frac{2}{3}$$

P(B_1 was chosen given a red ball and a green ball were selected)

$$= P(B_1/C) = \frac{P(B_1)P(C/B_1)}{P(B_1)P(C/B_1)+P(B_2)P(C/B_2)} = \frac{\frac{1}{2}\times\frac{2}{3}}{\left(\frac{1}{2}\times\frac{2}{3}\right)+\left(\frac{1}{2}\times\frac{2}{3}\right)} = \frac{1}{2}$$

(iii) Let D be the event that a green ball is selected.

$$P(D/B_1) == \frac{1}{3}, P(D/B_2) = \frac{2}{3}$$

$$= P(B_2/D) = \frac{P(B_2)P(D/B_2)}{P(B_1)P(D/B_1)+P(B_2)P(D/B_2)} = \frac{\frac{1}{2}\times\frac{2}{3}}{\left(\frac{1}{2}\times\frac{1}{3}\right)+\left(\frac{1}{2}\times\frac{2}{3}\right)} = \frac{2}{3}$$

(iv) Let E be the event that a white ball is selected.

The given two boxes does not contain a white ball, hence the probability is 0.

Random Variable:

A random variable is a real valued function whose domain is the sample space of a random experiment taking values on the real line $\mathbb{R}$.

Discrete Random Variable:

A discrete random variable is one which can take only finite or countable number of values with definite probabilities associated with each one of them.

Probability mass function:

Let X be discrete random variable which assuming values $x_1, x_2, \ldots, x_n$ with each of the values, we associate a number called the probability $P(X = x_i) = p(x_i), (i = 1, 2, \ldots, n)$ this is called the probability of x_i satisfying the following conditions

i. $p_i \geq 0 \ \forall i$ i.e., p_i's are all non-negative

ii. $\sum_{i=1}^{n} p_i = p_1 + p_2 + \ldots + p_n = 1$ i.e., the total probability is one.

Continuous random variable:

A continuous random variable is one which can assume every value between two specified values with a definite probability associated with each.

Probability Density Function:

A function f is said to be the probability density function of a continuous random variable X if it satisfies the following properties.

i. $f(x) \geq 0; \ -\infty < x < \infty$

ii. $\int_{-\infty}^{\infty} f(x)dx = 1.$

Distribution Function or Cumulative Distribution Function

i.Discrete Variable:

A distribution function of a discrete random variable X is defined as $P(X \leq x) = \sum_{x_i \leq x} P(x_i)$.

ii.Continuous Variable:

A distribution function of a continuous random variable X is defined as $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$.

Mathematical Expectation

The expected value of the random variable X is defined as i.If X is discrete random variable $E(X)=\sum_{i=1}^{\infty} x_i p(x_i)$ where $p(x)$ is the probability function of x.

ii.If X is continuous random variable $E(X)=\int_{-\infty}^{\infty} xf(x)dx$ where $f(x)$ is the probability density function of x.

Properties of Expectation:

1. If C is constant then $E(C)=C$

Proof:

Let X be a discrete random variable then $E(x)=\sum xp(x)$

Now $E(C)=\sum Cp(x)$

$=C\sum p(x)$ since $\sum_{i=1}^{n} p_i = p_1+p_2+\ldots+p_n=1$

$=C$

2. If a,b are constants then $E(ax+b)=aE(x)+b$

Proof:

Let X be a discrete random variable then $E(x)=\sum xp(x)$

Now $E(ax+b)=\sum(ax+b)p(x)$

$=\sum axp(x)+\sum bp(x)$

$=a\sum xp(x)+b\sum p(x)$ since $\sum_{i=1}^{n} p_i = p_1+p_2+\ldots+p_n=1$

$=aE(x)+b$

3. If a and b are constants then $Var(ax+b)=a^2Var(x)$

Proof:

$Var(ax+b)=E\left[\left(ax+b-E(ax+b)\right)^2\right]$

$=E\left[\left(ax+b-aE(x)-b\right)^2\right]$

$=E\left[a^2\left(x-E(x)\right)^2\right]$

$=a^2E\left[\left(x-E(x)\right)^2\right]$

$=a^2Var(x).$

4. If a is constant then $Var(ax)=a^2Var(x)$

Proof:

$Var(ax)=E\left[\left(ax-E(ax)\right)^2\right]$

$=E\left[\left(ax-aE(x)\right)^2\right]$

$=E\left[a^2\left(x-E(x)\right)^2\right]$

$=a^2E\left[\left(x-E(x)\right)^2\right]$

$=a^2Var(x).$

5. Prove that $Var(x)=E(x^2)-\left[E(x)\right]^2$

Proof:

$$Var(x)=E\left[\left(x-E\left(x\right)\right)^{2}\right]$$
$$=E\left[x^{2}+\left(E\left(x\right)\right)^{2}-2xE\left(x\right)\right]$$
$$=E\left[x^{2}+\mu^{2}-2x\mu\right]$$
$$=E\left(x^{2}\right)+E\left(\mu^{2}\right)-E\left(2x\mu\right)$$
$$=E\left(x^{2}\right)+\mu^{2}-2\mu E\left(x\right)$$
$$=E\left(x^{2}\right)+\mu^{2}-2\mu^{2}$$
$$=E\left(x^{2}\right)-\mu^{2}$$
$$Var(x)=E\left(x^{2}\right)-\left[E\left(x\right)\right]^{2}$$

Moment Generating Function (m.g.f)

A moment generating function of a random variable X (about origin) is defined as $M_X\left(t\right)=E\left(e^{tX}\right)=\begin{cases}\int e^{tx}f\left(x\right)dx \text{ , if x is continuous}\\ \sum e^{tx}p\left(x\right) \text{ , if x is discrete}\end{cases}$

Properties of Moment Generating Function

1. $M_{cx}\left(t\right)=M_x\left(ct\right)$

Proof:

$$M_{cx}\left(t\right)=E\left(e^{cxt}\right)$$
$$=E\left(e^{x(ct)}\right)$$
$$=M_x\left(ct\right)$$

2. $M_{x+c}\left(t\right)=e^{ct}M_x\left(t\right)$

Proof:

$$M_{x+c}\left(t\right)=E\left(e^{(x+c)t}\right)$$
$$=E(e^{xt}e^{ct})$$
$$=e^{ct}M_x\left(t\right)$$

3. $M_{ax+b}\left(t\right)=e^{bt}M_x\left(at\right)$

Proof:

$$M_{ax+b}\left(t\right)=E\left(e^{(ax+b)t}\right)$$
$$=E\left(e^{axt}e^{bt}\right)$$
$$=e^{bt}E\left(e^{x(at)}\right)$$
$$=e^{bt}M_x\left(at\right)$$

4. If X and Y are independent random variables then $M_{x+y}\left(t\right)=M_x\left(t\right).M_y\left(t\right)$

Proof:

$$M_{x+y}\left(t\right)=E\left(e^{(x+y)t}\right)$$
$$=E\left(e^{xt}e^{yt}\right)$$
$$=E\left(e^{xt}\right)E\left(e^{yt}\right)$$
$$M_{x+y}\left(t\right)=M_x\left(t\right)M_y\left(t\right)$$

Problem.1

If the probability distribution of X is given as

X	:	1	2	3	4
$P\ X$	:	0.4	0.3	0.2	0.1

Find $P(1/2 < X < 7/2/X > 1)$

Solution:

$$P\{1/2 < X < 7/2/X > 1\} = \frac{P\{(1/2 < X < 7/2) \cap X > 1\}}{P(X > 1)}$$

$$= \frac{P(X = 2 \, or \, 3)}{P(X = 2, 3 \, or \, 4)} \quad = \frac{P(X = 2) + P(X = 3)}{P(X = 2) + P(X = 3) + P(X = 4)}$$

$$= \frac{0.3 + 0.2}{0.3 + 0.2 + 0.1} = \frac{0.5}{0.6} = \frac{5}{6}.$$

Problem.2

A random variable X has the following probability distribution

X	:	-2	-1	0	1	2	3
$P\ X$	:	0.1	K	0.2	$2K$	0.3	$3K$

a) Find K, b) Evaluate $P(X < 2)$ and $P(-2 < X < 2)$

b) Find the cdf of X and d) Evaluate the mean of X.

Solution:

a) Since $\sum P(X) = 1$

$$0.1 + K + 0.2 + 2K + 0.3 + 3K = 1$$

$$6K + 0.6 = 1$$

$$6K = 0.4$$

$$K = \frac{0.4}{6} = \frac{1}{15}$$

b) $P(X < 2) = P(X = -2, -1, 0 \, or \, 1)$

$$= P(X = -2) + P(X = -1) + P(X = 0) + P(X = 1)$$

$$= \frac{1}{10} + \frac{1}{15} + \frac{1}{5} + \frac{2}{15} \quad = \frac{3 + 2 + 6 + 4}{30} = \frac{15}{30} = \frac{1}{2}$$

$P(-2 < X < 2) = P(X = -1, 0 \, or \, 1)$

$$= P(X = -1) + P(X = 0) + P(X = 1) = \frac{1}{15} + \frac{1}{5} + \frac{2}{15} \quad = \frac{1 + 3 + 2}{15} = \frac{6}{15} = \frac{2}{5}$$

c) The distribution function of X is given by $F(x)$ defined by

$X = x$	$P(X = x)$	$F(x) = P(X \le x)$
-2	$\frac{1}{10}$	$F(x) = P(X \le -2) = \frac{1}{10}$
-1	$\frac{1}{15}$	$F(x) = P(X \le -1) = \frac{1}{6}$
0	$\frac{2}{10}$	$F(x) = P(X \le 0) = \frac{11}{30}$
1	$\frac{2}{15}$	$F(x) = P(X \le 1) = \frac{1}{2}$
2	$\frac{3}{10}$	$F(x) = P(X \le 2) = \frac{4}{5}$
3	$\frac{3}{15}$	$F(x) = P(X \le 3) = 1$

d) Mean of X is defined by $E(X) = \sum xP(x)$

$$E(X) = \left(-2 \times \frac{1}{10}\right) + \left(-1 \times \frac{1}{15}\right) + \left(0 \times \frac{1}{5}\right) + \left(1 \times \frac{2}{15}\right) + \left(2 \times \frac{3}{10}\right) + \left(3 \times \frac{1}{5}\right)$$

$$= -\frac{1}{5} - \frac{1}{15} + \frac{2}{15} + \frac{3}{5} + \frac{3}{5} = \frac{16}{15}.$$

Problem.3

A random variable X has the following probability function:

X	:	0	1	2	3	4	5	6	7
$P\ X$	:	0	K	$2K$	$2K$	$3K$	K^2	$2K^2$	$7K^2+K$

Find (i) K , (ii) Evaluate $P(X<6), P(X \ge 6)$ and $P(0<X<5)$

(iii). Determine the distribution function of X .

(iv). $P(1.5 < X < 4.5/X > 2)$

(v). $E(3x-4), Var(3x-4)$

(vi). The smallest value of n for which $P(X \le n) > \frac{1}{2}$.

Solution:

(i) Since $\sum_{x=0}^{7} P(X) = 1,$

$$K + 2K + 2K + 3K + K^2 + 2K^2 + 7K^2 + K = 1$$

$$10K^2 + 9K - 1 = 0$$

$$K = \frac{1}{10} \quad or \quad K = -1$$

As $P(X)$ cannot be negative $K = \frac{1}{10}$

(ii) $P(X<6) = P(X=0) + P(X=1) + \ldots + P(X=5)$

$$= \frac{1}{10} + \frac{2}{10} + \frac{2}{10} + \frac{3}{10} + \frac{1}{100} + \ldots = \frac{81}{100}$$

Now $P(X \ge 6) = 1 - P(X<6)$

$$= 1 - \frac{81}{100} = \frac{19}{100}$$

Now $P(0<X<5) = P(X=1) + P(X=2) + P(X=3) = P(X=4)$

$$= K + 2K + 2K + 3K$$

$$= 8K = \frac{8}{10} = \frac{4}{5}.$$

(iii) The distribution of X is given by $F(x) = P(X \le x)$

$X = x$	$P(X=x)$	$F(x) = P(X \le x)$
0	0	$F(x) = P(X \le 0) = 0$
1	$\frac{1}{10}$	$F(x) = P(X \le 1) = \frac{1}{10}$
2	$\frac{2}{10}$	$F(x) = P(X \le 2) = \frac{3}{10}$
3	$\frac{2}{10}$	$F(x) = P(X \le 3) = \frac{5}{10}$
4	$\frac{3}{10}$	$F(x) = P(X \le 4) = \frac{8}{10}$
5	$\frac{1}{100}$	$F(x) = P(X \le 5) = \frac{81}{100}$
6	$\frac{2}{100}$	$F(x) = P(X \le 6) = \frac{83}{100}$
7	$\frac{17}{100}$	$F(x) = P(X \le 7) = 1$

(iv) $$P(1.5 < X < 4.5/X > 2) = \frac{P(x=3)+P(x=4)}{1-\left[P(x=0)+P(x=1)+P(x=2)\right]} = \frac{\frac{5}{10}}{1-\left[\frac{3}{10}\right]} = \frac{5}{7}$$

(v) $E(x) = \sum xp(x)$

$$= 1\times\frac{1}{10}+2\times\frac{2}{10}+3\times\frac{2}{10}+4\times\frac{3}{10}+5\times\frac{1}{100}+6\times\frac{2}{100}+7\times\frac{17}{100}$$

$E(x) = 3.66$

$E(x^2) = \sum x^2 p(x)$

$$= 1^2\times\frac{1}{10}+2^2\times\frac{2}{10}+3^2\times\frac{2}{10}+4^2\times\frac{3}{10}+5^2\times\frac{1}{100}+6^2\times\frac{2}{100}+7^2\times\frac{17}{100}$$

$E(x^2) = 16.8$

Mean $= E(x) = 3.66$

Variance $= E(x^2) - \left[E(x)\right]^2 = 16.8 - (3.66)^2 = 3.404$

(vi) The smallest value of n for which $P(X \le n) > \frac{1}{2}$ is 4

Problem.4

The probability mass function of random variable X is defined as $P(X=0) = 3C^2$, $P(X=1) = 4C - 10C^2$, $P(X=2) = 5C - 1$, where $C > 0$, and $P(X=r) = 0$ if $r \ne 0,1,2$. Find (i). The value of C.

(ii). $P(0 < X < 2/x > 0)$.

(iii). The distribution function of X.

(iv). The largest value of x for which $F(x) < \frac{1}{2}$.

Solution:

(i) Since $\sum_{x=0}^{x=2} p(x) = 1$

$p(0)+p(1)+p(2) = 1$

$3C^2 + 4C - 10C^2 + 5C - 1 = 1$

$7C^2 - 9C + 2 = 0$

$C = 1, \frac{2}{7}$

$C = 1$ is not applicable

$\therefore C = \frac{2}{7}$

The Probability distribution is

X	:	0	1	2
$P(X)$	:	$\frac{12}{49}$	$\frac{16}{49}$	$\frac{21}{49}$

(ii) $$P\left[0 < x < 2\big/ x > 0\right] = \frac{P\left[(0 < x < 2) \cap x > 0\right]}{P[x > 0]} = \frac{P[0 < x < 2]}{P[x > 0]}$$

$$= \frac{P[x=1]}{P[x=1]+P[X=2]}$$

$$P\left[0<x<2\big/ x>0\right]=\frac{\frac{16}{49}}{\frac{16}{49}+\frac{21}{49}}=\frac{16}{37}$$

(iii). The distribution function of X is

X	$F(X=x)=P(X\le x)$
0	$F(0)=P(X\le 0)=\frac{12}{49}=0.24$
1	$F(1)=P(X\le 1)=P(X=0)+P(X=1)=\frac{12}{49}+\frac{16}{49}=0.57$
2	$F(2)=P(X\le 2)=P(X=0)+P(X=1)+P(X=2)=\frac{12}{49}+\frac{16}{49}+\frac{21}{49}=1$

(iv) The Largest value of x for which $F(x)=P(X\le x)<\frac{1}{2}$ is 0.

Problem.5

If $P(x)=\begin{cases}\frac{x}{15}; x=1,2,3,4,5\\ 0 \quad ; elsewhere\end{cases}$

Find (i) $P\{X=1or2\}$ and (ii) $P\{1/2<X<5/2/x>1\}$

Solution:

i) $P(X=1\,or\,2)=P(X=1)+P(X=2)=\frac{1}{15}+\frac{2}{15}=\frac{3}{15}=\frac{1}{5}$

ii) $P\left(\frac{1}{2}<X<\frac{5}{2}/x>1\right)=\frac{P\left\{\left(\frac{1}{2}<X<\frac{5}{2}\right)\cap(X>1)\right\}}{P(X>1)}=\frac{P\{(X=1or2)\cap(X>1)\}}{P(X>1)}$

$=\frac{P(X=2)}{1-P(X=1)}$

$$=\frac{2/15}{1-(1/15)}=\frac{2/15}{14/15}=\frac{2}{14}=\frac{1}{7}.$$

Problem.6

A continuous random variable X has a probability density function $f(x)=3x^2$, $0\le x\le 1$. Find $'a'$ such that $P(X\le a)=P(X>a)$.

Solution:

Since $P(X\le a)=P(X>a)$, each must be equal to $\frac{1}{2}$ because the probability is always 1.

$\therefore P(X\le a)=\frac{1}{2}$

$\Rightarrow \int_0^a f(x)dx=\frac{1}{2}$

$\int_0^a 3x^2dx=\frac{1}{2}\Rightarrow 3\left[\frac{x^3}{3}\right]_0^a=a^3=\frac{1}{2}.$

$\therefore a=\left(\frac{1}{2}\right)^{\frac{1}{3}}$

Problem.7

A random variable X has the p.d.f $f(x)$ given by $f(x)=\begin{cases} Cxe^{-x}; \ if \ x>0 \\ 0 \quad ; \ if \ x\le 0 \end{cases}$ Find the value of C and cumulative density function of X.

Solution:

Since $\int_{-\infty}^{\infty} f(x)dx=1$

$$\int_{0}^{\infty} Cxe^{-x}dx=1$$

$$C\left[x\left(-e^{-x}\right)-\left(e^{-x}\right)\right]_0^{\infty}=1$$

$$C=1$$

$$\therefore f(x)=\begin{cases} xe^{-x}; x>0 \\ 0 \quad ; x\le 0 \end{cases}$$

Cumulative Distribution of x is

$$F(x)=\int_0^x f(x)dt=\int_0^x xe^{-x}dx=\left[-xe^{-x}-e^{-x}\right]_0^x=-xe^{-x}-e^{-x}+1$$

$$= 1-(1+x)e^{-x},\ x>0.$$

8. If a random variable X has the p.d.f $f(x)=\begin{cases} \frac{1}{2}(x+1); -1<x<1 \\ 0 \quad ; otherwise \end{cases}$. Find the mean and variance of X.

Solution:

$$\text{Mean}=\mu_1'=\int_{-1}^{1} xf(x)dx=\frac{1}{2}\int_{-1}^{1} x(x+1)dx=\frac{1}{2}\int_{-1}^{1}\left(x^2+x\right)dx=\frac{1}{2}\left(\frac{x^3}{3}+\frac{x^2}{2}\right)_{-1}^{1}=\frac{1}{3}$$

$$\mu_2'=\int_{-1}^{1} x^2 f(x)dx=\frac{1}{2}\int_{-1}^{1}\left(x^3+x^2\right)dx=\frac{1}{2}\left[\frac{x^4}{4}+\frac{x^3}{3}\right]_{-1}^{1}=\frac{1}{2}\left[\frac{1}{4}+\frac{1}{3}-\frac{1}{4}+\frac{1}{3}\right]=\frac{1}{2}.\frac{2}{3}=\frac{1}{3}$$

$$Variance=\mu_2'-\left(\mu_1'\right)^2=\frac{1}{3}-\frac{1}{9}=\frac{3-1}{9}=\frac{2}{9}.$$

9. A continuous random variable X that can assume any value between $X=2$ and $X=5$ has a probability density function given by $f(x)=k(1+x)$. Find $P(X<4)$.

Solution:

Given X is a continuous random variable whose pdf is $f(x)=\begin{cases} k(1+x), 2<x<5 \\ 0 \ , Otherwise \end{cases}$.

Since $\int_{-\infty}^{\infty} f(x)dx=1 \Rightarrow \int_2^5 k(1+x)dx=1$

$$k\left[\frac{(1+x)^2}{2}\right]_2^5=1$$

$$k\left[\frac{(1+5)^2}{2}-\frac{(1+2)^2}{2}\right]=1$$

$$k\left[18-\frac{9}{2}\right]=1$$

$$k\left[\frac{27}{2}\right]=1\Rightarrow k=\frac{2}{27}$$

$$\therefore f(x)=\begin{cases}\dfrac{2(1+x)}{27}, 2<x<5\\ 0 \quad , Otherwise\end{cases}$$

$$P(X<4)=\frac{2}{27}\int_2^4 (1+x)\,dx$$

$$=\frac{2}{27}\left[\frac{(1+x)^2}{2}\right]_2^4=\frac{2}{27}\left[\frac{(1+4)^2}{2}-\frac{(1+2)^2}{2}\right]=\frac{2}{27}\left[\frac{25}{2}-\frac{9}{2}\right]=\frac{2}{27}\frac{16}{2}=\frac{16}{27}.$$

10. A random variable X has density function given by $f(x)=\begin{cases}2e^{-2x}; x\geq 0\\ 0 \quad ; x<0\end{cases}$. Find m.g.f

Solution:

$$M_X(t)=E\left(e^{tx}\right)=\int_0^\infty e^{tx} f(x)\,dx=\int_0^\infty e^{tx}2e^{-2x}dx$$

$$=2\int_0^\infty e^{(t-2)x}dx$$

$$=2\left[\frac{e^{(t-2)x}}{t-2}\right]_0^\infty=\frac{2}{2-t}, t<2.$$

11. The pdf of a random variable X is given by $f(x)=\begin{cases}2x, \ 0\leq x\leq b\\ 0, \ \ otherwise\end{cases}$. For what value of b is $f(x)$ a valid pdf? Also find the cdf of the random variable X with the above pdf.

Solution:

Given $f(x)=\begin{cases}2x, \ 0\leq x\leq b\\ 0, \ \ otherwise\end{cases}$

Since $\int_{-\infty}^{\infty} f(x)\,dx=1 \Rightarrow \int_0^b 2x\,dx=1$

$$\left[2\frac{x^2}{2}\right]_0^b=1$$

$$\left[b^2-0\right]=1 \ \Rightarrow \text{ b =1}$$

$$\therefore f(x)=\begin{cases}2x, \ 0\leq x\leq 1\\ 0, \ \ otherwise\end{cases}$$

$$F(x)=P(X\leq x)=\int_0^x f(x)\,dx=\int_0^x 2xdx=\left[2\frac{x^2}{2}\right]_0^x=x^2 \ , 0\leq x\leq 1$$

$$F(x)=P(X\leq x)=\int_{-\infty}^x f(x)\,dx=\int_{-\infty}^x 0\,dx=0 \ , \ x<0$$

$$F(x)=P(X\leq x)=\int_{-\infty}^0 f(x)\,dx+\int_0^1 f(x)\,dx+\int_1^x f(x)\,dx \qquad =\int_{-\infty}^0 0\ dx+\int_0^1 2x\,dx+\int_1^x 0\,dx =$$

$$\left[2\frac{x^2}{2}\right]_0^1=1 \ , \ x>1$$

$$F(x)=\begin{cases}0, & x<0\\ x^2, & 0\leq x\leq 1\\ 1, & x>1\end{cases}$$

12. A random variable X has density function $f(x)=\begin{cases}\dfrac{K}{1+x^2}, -\infty<x<\infty\\ 0 \quad , \ Otherwise\end{cases}$.

Determine K and the distribution functions. Evaluate the probability $P(x\geq 0)$.

Solution:

$$\text{Since } \int_{-\infty}^{\infty} f(x)dx=1$$

$$\int_{-\infty}^{\infty}\frac{K}{1+x^2}dx=1$$

$$K\int_{\infty}^{\infty}\frac{dx}{1+x^2}=1$$

$$K\left(\tan^{-1}x\right)_{-\infty}^{\infty}=1$$

$$K\left(\frac{\pi}{2}-\left(-\frac{\pi}{2}\right)\right)=1$$

$$K\pi=1$$

$$K=\frac{1}{\pi}$$

$$F(x)=\int_{-\infty}^{x} f(x)dx=\int_{-\infty}^{x}\frac{K}{1+x^2}dx$$

$$=\frac{1}{\pi}\left[\tan^{-1}x-\left(-\frac{\pi}{2}\right)\right]$$

$$F(x)=\frac{1}{\pi}\left[\frac{\pi}{2}+\tan^{-1}x\right], -\infty<x<\infty$$

$$P(X\geq 0)=\frac{1}{\pi}\int_{0}^{\infty}\frac{dx}{1+x^2}=\frac{1}{\pi}\left(\tan^{-1}x\right)_{0}^{\infty}$$

$$=\frac{1}{\pi}\left(\frac{\pi}{2}-\tan^{-1}0\right)=\frac{1}{2}.$$

13. If X has the probability density function $f(x)=\begin{cases}Ke^{-3x}, x>0\\ 0 \quad ,otherwise\end{cases}$ find K, $P[0.5\leq X\leq 1]$ and the mean of X.

Solution:

$$\text{Since } \int_{-\infty}^{\infty} f(x)\,dx=1$$

$$\int_{0}^{\infty}Ke^{-3x}dx=1$$

$$K\left[\frac{e^{-3x}}{-3}\right]_{0}^{\infty}=1$$

$$\frac{K}{3}=1$$

$$K=3$$

$$P(0.5\leq X\leq 1)=\int_{0.5}^{1} f(x)\,dx=3\int_{0.5}^{1}e^{-3x}dx=\cancel{3}\left(\frac{e^{-3}-e^{-1.5}}{-\cancel{3}}\right)=\left[e^{-1.5}-e^{-3}\right]$$

Mean of $X = E(x) = \int_0^\infty xf(x)\,dx = 3\int_0^\infty xe^{-3x}dx$

$$= 3\left[x\left(\frac{-e^{-3x}}{3}\right) - 1\left(\frac{e^{-3x}}{9}\right)\right]_0^\infty = \frac{3\times 1}{9} = \frac{1}{3}$$

Hence the mean of $X = E(X) = \frac{1}{3}$.

14. If X is a continuous random variable with pdf given by $f(x) = \begin{cases} Kx & in \quad 0 \le x \le 2 \\ 2K & in \quad 2 \le x \le 4 \\ 6K - Kx & in \quad 4 \le x \le 6 \\ 0 & elsewhere \end{cases}$. Find the value of K and also the cdf $F(x)$.

Solution:

$$\text{Since } \int_{\infty}^{\infty} F(x)\,dx = 1$$

$$\int_0^2 Kxdx + \int_2^4 2Kdx + \int_4^6 (6k - kx)dx = 1$$

$$K\left[\left(\frac{x^2}{2}\right)_0^2 + (2x)_2^4 + \int_4^6 \left(6x - \frac{x^2}{2}\right)_4^6\right] = 1$$

$$K[\cancel{2} + \cancel{8} - 4 + 36 - 18 - 24 + 8] = 1$$

$$8K = 1$$

$$K = \frac{1}{8}$$

We know that $F(x) = \int_{-\infty}^{x} f(x)\,dx$

If $x < 0$, then $F(x) = \int_{-\infty}^{x} f(x)\,dx = 0$

If $x \in (0,2)$, then $F(x) = \int_{-\infty}^{x} f(x)\,dx$

$$F(x) = \int_{-\infty}^{0} f(x)\,dx + \int_0^x f(x)\,dx$$

$$= \int_{-\infty}^{0} 0dx + \int_0^x Kxdx = \int_{-\infty}^{0} 0dx + \frac{1}{8}\int_0^x xdx$$

$$F(x) = \left(\frac{x^2}{16}\right)_0^x = \frac{x^2}{16}, 0 \le x \le 2$$

If $x \in (2,4)$, then $F(x) = \int_{-\infty}^{0} f(x)\,dx + \int_0^2 f(x)\,dx + \int_2^x f(x)\,dx$

$$= \int_{-\infty}^{0} 0dx + \int_0^2 Kxdx + \int_2^x 2Kdx$$

$$= \int_0^2 \frac{x}{8}dx + \int_2^x \frac{1}{4}dx = \left(\frac{x^2}{16}\right)_0^2 + \left(\frac{x}{4}\right)_2^x$$

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

$$F(x)=\frac{x}{4}-\frac{4}{16}=\frac{x-1}{4},\ 2\le x<4$$

If $x\in(4,6)$, then $F(x)=\int_{-\infty}^{0}0dx+\int_{0}^{2}Kxdx+\int_{2}^{4}2Kdx+\int_{4}^{x}K(6-x)dx$

$$=\int_{0}^{2}\frac{x}{8}dx+\int_{2}^{4}\frac{1}{4}dx+\int_{4}^{x}\frac{1}{8}(6-x)dx$$

$$=\left(\frac{x^2}{16}\right)_0^2+\left(\frac{x}{4}\right)_2^4+\left(\frac{6x}{8}-\frac{x^2}{16}\right)_4^x$$

$$=\frac{1}{4}+1-\frac{1}{2}+\frac{6x}{8}-\frac{x^2}{16}-3+1$$

$$=\frac{4+16-8+12x-x^2-48+16}{16}$$

$$F(x)=\frac{-x^2+12x-20}{16},4\le x\le 6$$

If $x>6$, then $F(x)=\int_{-\infty}^{0}0dx+\int_{0}^{2}Kxdx+\int_{2}^{4}2Kdx+\int_{4}^{6}K(6-x)dx+\int_{6}^{\infty}0dx$

$$F(x)=1,\ x\ge 6$$

$$\therefore F(x)=\begin{cases}0 & ;x\le 0\\ \frac{x^2}{16} & ;0\le x\le 2\\ \frac{1}{4}(x-1) & ;2\le x\le 4\\ \frac{-1}{16}\left(20-12x+x^2\right) & ;4\le x\le 6\\ \quad 1 & ;x\ge 6\end{cases}$$

15. A random variable X has the P.d.f $f(x)=\begin{cases}2x,0<x<1\\ 0\ \ ,Otherwise\end{cases}$

Find (i) $P\left(X<\frac{1}{2}\right)$ (ii) $P\left(\frac{1}{4}<x<\frac{1}{2}\right)$ (iii) $P\left(X>\frac{3}{4}/X>\frac{1}{2}\right)$

Solution:

(i) $P\left(x<\frac{1}{2}\right)=\int_{0}^{1/2}f(x)dx=\int_{0}^{1/2}2xdx=2\left(\frac{x^2}{2}\right)_0^{1/2}=\frac{2\times 1}{8}=\frac{1}{4}$

(ii) $P\left(\frac{1}{4}<x<\frac{1}{2}\right)=\int_{1/4}^{1/2}f(x)dx=\int_{1/4}^{1/2}2xdx=2\left(\frac{x^2}{2}\right)_{1/4}^{1/2}$

$$=2\left(\frac{1}{8}-\frac{1}{32}\right)=\left(\frac{1}{4}-\frac{1}{16}\right)=\frac{3}{16}.$$

(iii) $P\left(X>\frac{3}{4}/X>\frac{1}{2}\right)=\frac{P\left(X>\frac{3}{4}\cap X>\frac{1}{2}\right)}{P\left(X>\frac{1}{2}\right)}=\frac{P\left(X>\frac{3}{4}\right)}{P\left(X>\frac{1}{2}\right)}$

$$P\left(X>\frac{3}{4}\right)=\int_{3/4}^{1}f(x)dx=\int_{3/4}^{1}2xdx=2\left(\frac{x^2}{2}\right)_{3/4}^{1}=1-\frac{9}{16}=\frac{7}{16}$$

$$P\left(X>\frac{1}{2}\right)=\int_{1/2}^{1}f(x)dx=\int_{1/2}^{1}2xdx=2\left(\frac{x^2}{2}\right)_{1/2}^{1}=1-\frac{1}{4}=\frac{3}{4}$$

$$P\left(X>\frac{3}{4}/X>\frac{1}{2}\right)=\frac{\frac{7}{16}}{\frac{3}{4}}=\frac{7}{16}\times\frac{4}{3}=\frac{7}{12}.$$

16. Let the random variable X have the p.d.f $f(x)=\begin{cases}\frac{1}{2}e^{-\frac{x}{2}} & ,x>0\\ 0 & ,otherwise.\end{cases}$ .Find the moment generating function, mean & variance of X .

Solution:

$$M_X(t)=E\left(e^{tx}\right)=\int_{-\infty}^{\infty}e^{tx}f(x)dx=\int_{0}^{\infty}e^{tx}\frac{1}{2}e^{-x/2}dx$$

$$=\frac{1}{2}\int_0^{\infty}e^{-\left(\frac{1}{2}-t\right)x}dx=\frac{1}{2}\left[\frac{e^{-\left(\frac{1}{2}-t\right)x}}{-\left(\frac{1}{2}-t\right)}\right]_0^{\infty}=\frac{1}{1-2t},\ if\ t<\frac{1}{2}.$$

$$E(X)=\left[\frac{d}{dt}M_X(t)\right]_{t=0}=\left[\frac{2}{(1-2t)^2}\right]_{t=0}=2$$

$$E\left(X^2\right)=\left[\frac{d^2}{dt^2}M_X(t)\right]_{t=0}=\left[\frac{8}{(1-2t)^3}\right]_{t=0}=8$$

$$Var(X)=E\left(X^2\right)-\left[E(X)\right]^2=8-4=4.$$

17. The first four moments of a distribution about $x=4$ are 1,4,10 and 45 respectively. Show that the mean is 5, variance is 3, $\mu_3=0$ and $\mu_4=26$.

Solution:

Given $\mu_1'=1, \mu_2'=4, \mu_3'=10, \mu_4'=45$

$\mu_r'=r^{th}$ moment about to value $x=4$

Here $A=4$

Here Mean $=A+\mu_1'=4+1=5$

$$\text{Variance}=\mu_2=\mu_2'-\left(\mu_1'\right)^2$$

$$=4-1=3.$$

$$\mu_3=\mu_3'-3\mu_2'\mu_1'+2\left(\mu_1'\right)^3$$

$$=10-3(4)(1)+2(1)^3=0$$

$$\mu_4=\mu_4'-4\mu_3'\mu_1'+6\mu_2'\left(\mu_1'\right)^2-3\left(\mu_1'\right)^4$$

$$=45-4(10)(1)+6(4)(1)^2-3(1)^4$$

$$\mu_4=26.$$

18. Find the moment generating function and rth moments for the distribution. Whose p.d.f is $f(x)=Ke^{-x},\ 0\le x\le\infty$. Find also standard deviation.

Solution:

Total Probability=1

$$\therefore\int_0^{\infty}ke^{-x}dx=1$$

$$k\left[\frac{e^{-x}}{-1}\right]_0^{\infty}=1$$
$$k=1$$

$$M_X(t)=E\left[e^{tx}\right]=\int_0^{\infty}e^{tx}e^{-x}dx=\int_0^{\infty}e^{(t-1)x}dx$$

$$=\left[\frac{e^{(t-1)x}}{t-1}\right]_0^{\infty}=\frac{1}{1-t},t<1$$

$$=(1-t)^{-1}=1+t+t^2+\ldots+t^r+\ldots\infty$$

$$\mu_r'=coeff.\ of\ \frac{t^r}{r!}=r!$$

When $r=1,\ \mu_1'=1!=1$

$r=2,\mu_2'=2!=2$

Variance$=\mu_2'-\mu_1'=2-1=1$

$\therefore$ Standard deviation=1.

19. A continuous random variable X has the p.d.f $f(x)=kx^2e^{-x},\ x\geq 0$. Find the rth moment of X about the origin. Hence find mean and variance of X.

Solution:

Since $\int_0^{\infty}Kx^2e^{-x}dx=1$

$$K\left[x^2\left(\frac{e^{-x}}{-1}\right)-2x\left(\frac{e^{-x}}{1}\right)+2\left(\frac{e^{-x}}{-1}\right)\right]_0^{\infty}=1$$
$$2K=1$$
$$K=\frac{1}{2}.$$

$$\mu_r'=\int_0^{\infty}x^r f(x)dx$$

$$=\frac{1}{2}\int_0^{\infty}x^{r+2}e^{-x}dx$$

$$=\frac{1}{2}\int_0^{\infty}e^{-x}x^{(r+3)-1}dx=\frac{(r+2)!}{2}$$

Putting $n=1,\mu_1'=\frac{3!}{2}=3$

$n=2,\mu_2'=\frac{4!}{2}=12$

$\therefore$ Mean = $\mu_1'=3$

Variable= $\mu_2'-\left(\mu_1'\right)^2$

i.e. $\mu_2=12-(3)^2=12-9$

$\therefore\mu_2=3.$

20. Find the moment generating function of the random variable X, with probability density function $f(x)=\begin{cases}x & for\quad 0\leq x<1\\ 2-x & for\quad 1\leq x<2\\ 0 & otherwise\end{cases}$.Also find μ_1',μ_2'.

Solution:

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

$$=\int_0^1 e^{tx}xdx+\int_1^2 e^{tx}(2-x)dx$$

$$=\left(\frac{xe^{tx}}{t}-\frac{e^{tx}}{t^2}\right)_0^1+\left[(2-x)\frac{e^{tx}}{t}-(-1)\frac{e^{tx}}{t^2}\right]_1^2$$

$$=\frac{e^t}{t}-\frac{e^t}{t^2}+\frac{1}{t^2}+\frac{e^{2t}}{t^2}-\frac{e^t}{t}-\frac{e^t}{t^2}$$

$$=\left(\frac{e^t-1}{t}\right)^2$$

$$=\left[1+\frac{t}{1!}+\frac{t^2}{2!}+\frac{t^3}{3!}+\ldots-1\right]^2$$

$$=\left[1+\frac{t}{2!}+\frac{t^2}{3!}+\frac{t^3}{4!}+\ldots\right]^2$$

$$\mu_1'=coeff.\ of\ \frac{t}{1!}=1$$

$$\mu_2'=coeff.\ of\ \frac{t^2}{2!}=\frac{7}{6}.$$

21. The p.d.f of the r.v. X follows the probability law: $f(x)=\frac{1}{2\theta}e^{-\frac{|x-\theta|}{\theta}}$, $-\infty<x<\infty$. Find the m.g.f of X and also find $E\ \ X$ and $V\ \ X$.

Solution:

$$M_X(t)=E\left(e^{tx}\right)=\int_{-\infty}^{\infty} e^{tx} f(x)dx=\int_{-\infty}^{\infty}\frac{1}{2\theta}e^{\frac{-|x-\theta|}{\theta}}e^{tx}dx$$

$$=\int_{-\infty}^{\theta}\frac{1}{2\theta}e^{\frac{(x-\theta)}{\theta}}e^{tx}dx+\int_{\theta}^{\infty}\frac{1}{2\theta}e^{\frac{-(x-\theta)}{\theta}}e^{tx}dx$$

$$M_X(t)=\frac{e^{-1}}{2\theta}\int_{-\infty}^{\theta}e^{x\left(t+\frac{1}{\theta}\right)}dx+\frac{e}{2\theta}\int_{\theta}^{\infty}e^{-x\left(\frac{1}{\theta}-t\right)}dx$$

$$=\frac{e^{-1}}{2\theta}\frac{e^{\theta\left(t+\frac{1}{\theta}\right)}}{\left(t+\frac{1}{\theta}\right)}+\frac{e}{2\theta}\frac{e^{-\theta\left(\frac{1}{\theta}-1\right)}}{\left(\frac{1}{\theta}-t\right)}$$

$$=\frac{e^{\theta t}}{2(\theta t+1)}+\frac{e^{\theta t}}{2(1-\theta t)}=\frac{e^{\theta t}}{1-\theta^2t^2}=e^{\theta t}\left[1-(\theta t)^2\right]^{-1}$$

$$=\left[1+\theta t+\frac{\theta^2t^2}{2!}+\ldots\right]\left[1+\theta^2t^2+\theta^4t^4+\ldots\right]$$

$$=1+\theta t+\frac{3\theta^2t^2}{2!}+\ldots$$

$$E(X)=\mu_1'=coeff.\ of\ t\ in\ M_X(t)=\theta$$

$$\mu_2'=coeff.\ of\ \frac{t^2}{2!}\ in\ M_X(t)=3\theta^2$$

$$Var(X)=\mu_2'-\left(\mu_1'\right)^2=3\theta^2-\theta^2=2\theta^2.$$

22. The elementary probability law of a continues random variable is

$f(x)=y_0e^{-b(x-a)}$, $a\le x\le\infty, b>0$ where a, b and y_0 are constants. Find y_0 the r^th^ moment about point $x=a$ and also find the mean and variance.

Solution:

Since the total probability is unity,

$$\int_{-\infty}^{\infty} f(x)\,dx=1$$

$$y_0\int_0^{\infty} e^{-b(x-a)}dx=1$$

$$y_0\left[\frac{e^{-b(x-a)}}{-b}\right]_0^{\infty}=1$$

$$y_0\left(\frac{1}{b}\right)=1$$

$$y_0=b.$$

μ_r' (r^th^ moment about the point $x=a$) $=\int_{-\infty}^{\infty}(x-a)^r f(x)\,dx$

$$=b\int_a^{\infty}(x-a)^r e^{-b(x-a)}dx$$

Put $x-a=t$, $dx=dt$, when $x=a, t=0$, $x=\infty, t=\infty$

$$=b\int_0^{\infty} t^r e^{-bt}dt$$

$$=b\frac{\Gamma(r+1)}{b^{(r+1)}}=\frac{r!}{b^r}$$

In particular $r=1$

$$\mu_1'=\frac{1}{b}$$

$$\mu_2'=\frac{2}{b^2}$$

Mean $=a+\mu_1'=a+\dfrac{1}{b}$

Variance $=\mu_2'-\left(\mu_1'\right)^2$

$$=\frac{2}{b^2}-\frac{1}{b^2}=\frac{1}{b^2}.$$

23. The lifetime (in hours) of a certain piece of equipment is a continuous r.v. having range $0<x<\infty$ and p.d.f.is $f(x)=\begin{cases} xe^{-kx}, 0<x<\infty \\ 0 \quad ,otherwise\end{cases}$. Determine the constant K and evaluate the probability that the life time exceeds 2 hours.

Solution:

Let X the life time of a certain piece of equipment.

Then the p.d.f. $f(x)=\begin{cases} xe^{-kx}, 0<x<\infty \\ 0 \quad ,Otherwise\end{cases}$

To find K, $\int_0^{\infty} f(x)\,dx=1$

$$\int_0^{\infty} e^{-kx}x^{2-1}dx=1$$

$$\frac{\Gamma(2)}{K^2}=1\Rightarrow K^2=1\ \therefore K=1$$

$$\therefore f(x)=\begin{cases} xe^{-x}, 0<x<\infty \\ 0 \quad ,Otherwise\end{cases}$$

P[Life time exceeds 2 hours] $= P[X>2]$

$$=\int_2^\infty f(x)dx$$

$$=\int_2^\infty xe^{-x}dx$$

$$=\left[x\left(-e^{-x}\right)-\left(e^{-x}\right)\right]_2^\infty$$

$$=2e^{-2}+e^{-2}=3e^{-2}=0.4060$$

24. If the continuous random variable X has ray Leigh density $F(x)=\left(\frac{x}{\alpha^2}e^{-\frac{x^2}{2\alpha^2}}\right)\times U(x)$ find $E\left(x^n\right)$and deduce the values of $E(X)$ and $Var(X)$.

Solution:

Here $U(x)=\begin{cases}1 & if \quad x>0 \\ 0 & if \quad x\le 0\end{cases}$

$$E\left(x^n\right)=\int_0^\infty x^n f(x)dx$$

$$=\int_0^\infty x^n \frac{x}{\alpha^2}e^{\frac{-x^2}{2\alpha^2}}dx$$

Put $\frac{x^2}{2\alpha^2}=t,$ $\qquad x=0,\ t=0$

$\frac{x}{\alpha^2}dx=dt$ $\qquad x=\alpha, t=\infty$

$$=\int_0^\infty \left(2\alpha^2 t\right)^{n/2}e^{-t}dt \quad \left[\because x=\sqrt{2\alpha}.\sqrt{t}\right]$$

$$=2^{n/2}\alpha^n\int_0^\infty t^{n/2}e^{-t}dt$$

$$E\left(x^n\right)=2^{n/2}\alpha^n\Gamma\left(\frac{n}{2}+1\right)-(1)$$

Putting $n=1$ in (1) we get

$$E(x)=2^{1/2}\alpha\,\Gamma\left(\frac{3}{2}\right)=\sqrt{2}\alpha\,\Gamma\left(\frac{1}{2}+1\right)$$

$$=\sqrt{2}\alpha\frac{1}{2}\Gamma\left(\frac{1}{2}\right)$$

$$=\frac{\alpha}{\sqrt{2}}\sqrt{\pi} \quad [\because \Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$$

$$\therefore E(x)=\alpha\sqrt{\frac{\pi}{2}}$$

Putting $n=2$ in (1), we get

$$E\left(x^2\right)=2\alpha^2\Gamma(2)=2\alpha^2 \quad [\because \Gamma(2)=1]$$

$$\therefore Var(X)=E\left(X^2\right)-\left[E(X)\right]^2$$

$$=2\alpha^2-\alpha^2\frac{\pi}{2}$$

$$=\left(2-\frac{\pi}{2}\right)\alpha^2=\left(\frac{4-\pi}{2}\right)\alpha^2.$$

Standard Distributions

Discrete type

Binomial distribution:

A random variable X is said to follow binomial distribution if it assumes only non negative values and its probability mass function is given by $P(X=x)=p(x)=\begin{cases} nC_x p^x q^{n-x}, x=0,1,2,\ldots,n; q=1-p \\ 0, otherwise \end{cases}$

Notation: $X \sim B(n,p)$ read as X is following binomial distribution with parameter n and p.

1. Find m.g.f. of Binomial distribution and find its mean and variance.

Solution:

M.G.F.of Binomial distribution:-

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

$$=\sum_{x=0}^{n} nC_x x\, P^x q^{n-x}e^{tx}$$

$$=\sum_{x=0}^{n} nC_x\left(pe^t\right)^x q^{n-x}$$

$$M_X(t)=\left(q+pe^t\right)^n$$

Mean of Binomial distribution

Mean $=E(X)=M_X{}'(0)$

$$=\left[n\left(q+pe^t\right)^{n-1}pe^t\right]_{t=0}=np \text{ Since } q+p=1$$

$$E\left(X^2\right)=M_X{}''(0)$$

$$=\left[n(n-1)\left(q+pe^t\right)^{n-2}\left(pe^t\right)^2+npe^t\left(q+pe^t\right)^{n-1}\right]_{t=0}$$

$$E\left(X^2\right)=n(n-1)p^2+np$$

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

Variance $=E\left(X^2\right)-\left[E[X]\right]^2=npq$

$$Mean=np\,; Variance=npq$$

2. Comment the following: “The mean of a binomial distribution is 3 and variance is 4

Solution:

In binomial distribution, mean $>$ variance but Variance $<$ Mean

Since Variance $=4$ &Mean $=3$, the given statement is wrong.

3. If X and Y are independent binomial variates $B\left(5,\frac{1}{2}\right)$ and $B\left(7,\frac{1}{2}\right)$ find $P[X+Y=3]$

Solution:

$X+Y$ is also a binomial variate with parameters $n_1+n_2=12$ & $p=\frac{1}{2}$

$$\therefore P[X+Y=3]=12C_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^9=\frac{55}{2^{10}}$$

4. (i). Six dice are thrown 729 times. How many times do you expect atleast 3 dice show 5 or 6 ?

(ii) Six coins are tossed 6400 times. Using the Poisson distribution, what is the approximate probability of getting six heads 10 times?

Solution:

(i). Let X be the number of times the dice shown 5 or 6

$$P[5\,or\,6] = \frac{1}{6}+\frac{1}{6}=\frac{1}{3}$$

$$\therefore P=\frac{1}{3} \text{ and } q=\frac{2}{3}$$

Here $n=6$

By Binomial theorem,

$$P[X=x]=6C_x\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{6-x} \text{ where } x=0,1,2\ldots6.$$

$$P[X\geq 3]=P(3)+P(4)+P(5)+P(6)$$

$$=6C_3\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^3+6C_4\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^2+6C_5\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)+6C_6\left(\frac{1}{3}\right)^6$$

$$=0.3196$$

$\therefore$ Expected number of times atleast 3 dies to show 5 or 6 $=N\times P[X\geq 3]$

$$=729\times 0.3196=233.$$

(ii). Probability of getting six heads in one toss of six coins is $p=\left(\frac{1}{2}\right)^6$,

$$\lambda=np=6400\times\left(\frac{1}{2}\right)^6=100$$

Let X be the number of times getting 6 heads $P(X=10)=\dfrac{e^{-100}(100)^{10}}{10!}=1.025\times 10^{-30}$

Poisson distribution:

A random variable X is said to follow Poisson distribution if it assumes only non negative values and its probability mass function is given by

$$P(X=x)=\begin{cases}\dfrac{e^{-\lambda}\lambda^x}{x!};x=0,1,2,\ldots;\lambda>0\\ 0,otherwise\end{cases}$$

Notation: $X\sim P(\lambda)$ read as X is following Poisson distribution with parameter λ.

Poisson distribution as limiting form of binomial distribution:

Poisson distribution is a limiting case of Binomial distribution under the following conditions:

(i). n the number of trials is indefinitely large, (i.e.) $n\to\infty$

(ii). p the constant probability of success in each trial is very small (i.e.) $p\to 0$

(iii). $np=\lambda$ is finite.

Proof:

$$P(X=x)=p(x)=nc_x\,p^x\,q^{n-x}$$

Let $np=\lambda$

$$\therefore \qquad p=\frac{\lambda}{n},\; q=1-\frac{\lambda}{n}$$

$$\therefore \qquad p(x)=nc_x\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$=\frac{n!}{x!\,(n-x)!}\left(\frac{\lambda}{n}\right)^{x}\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$=\frac{n(n-1)\cdots(n-(x-1))\,\cancel{(n-x)!}}{x!\,\cancel{(n-x)!}}\left(\frac{\lambda}{n}\right)^{x}\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$=\frac{1.\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)}{x!}\cancel{n^{x}}\,\frac{\lambda^{x}}{\cancel{n^{x}}}\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$p(x)=1.\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\frac{\lambda^{x}}{x!}\left(1-\frac{\lambda}{n}\right)^{n-x}$$

Taking limit $n\to\infty$ on both sides

$$\lim_{n\to\infty}p(x)=\frac{\lambda^{x}}{x!}\lim_{n\to\infty}\left[\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{n-x}\right]$$

$$=\frac{\lambda^{x}}{x!}\lim_{n\to\infty}\left[\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\right]\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}$$

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

Problem.1 Criticise the following statement: “The mean of a Poisson distribution is 5 while the standard deviation is 4”.

Solution:

For a Poisson distribution mean and variance are same. Hence this statement is not true.

Problem.2 If *X* is a Poisson variate $P(X=2)=9P(X=4)+90P(X=6)$, find mean and variance of *X*.

Solution:

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

$$P(X=2)=9P(X=4)+90P(X=6)$$

$$\frac{e^{-\lambda}\lambda^{2}}{2!}=9\frac{e^{-\lambda}\lambda^{4}}{4!}+90\frac{e^{-\lambda}\lambda^{6}}{6!}$$

$$\frac{1}{2!}=\frac{9\lambda^{2}}{4!}+90\frac{\lambda^{4}}{6!}$$

$$\frac{1}{2}=\frac{3}{8}\lambda^{2}+\frac{\lambda^{4}}{8}$$

$$1=\frac{3\lambda^{2}}{4}+\frac{\lambda^{4}}{4}$$

$$\lambda^{4}+3\lambda^{3}-4=0$$

Put $\lambda^{2}=t$, $\quad t^{2}+3t-4=0$

$$(t+4)(t-1)=0$$

$$t=1,-4$$

$\therefore\quad \lambda^{2}=1,\qquad \lambda^{2}=-4$

$\lambda=\pm1,\qquad \lambda=\pm2i$

$\therefore\quad$ Mean $=\lambda=1\ \ (\because \lambda>0)$

Variance $=\lambda=1.$

Problem.3 If X is a Poisson rv such that $P(X=1)=0.3$ and $P(X=2)=0.2$. Find $P(X=0)$.

Solution:

Given X is a Poisson rv, $p(X=x)=e^{-\lambda}\frac{\lambda^x}{x!},\ x=0,1,\ldots$

$$P(X=1)=\frac{e^{-\lambda}\lambda^1}{1!}=0.3 \qquad (1)$$

$$P(X=2)=\frac{e^{-\lambda}\lambda^2}{2!}=0.2 \qquad (2)$$

$$\frac{(1)}{(2)}\Rightarrow \qquad \frac{e^{-\lambda}\lambda^1}{e^{-\lambda}\lambda^2}2=\frac{0.3}{0.2}$$

$$\frac{1}{\lambda}=\frac{0.3}{2(0.2)}$$

$$\lambda=1.3333$$

$$\therefore \qquad P(X=x)=e^{-1.3333}\frac{(1.3333)^x}{x!}$$

$$\therefore \qquad P(X=0)=e^{-1.3333}\frac{(1)}{1}=0.2636.$$

Problem.4 Out of 800 families with 4 children each, how many families would be expected to have (i) 2 boys and 2 girls, (ii) at least 1 boy, (iii) at most 2 girls and (iv) children of both sexes. Assume equal probabilities for boys and girls.

Solution:

Considering each child as a trial, $n=4$. Assuming that birth of a boy is a success,

$$p=\frac{1}{2} \qquad \therefore\ q=\frac{1}{2}.$$

Let X denote the number of successes (boys)

(i) $P(\text{2 boys and 2 girls})=P(X=2)$

$$=4c_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2$$

$$=0.375$$

$\therefore$ No. of families having 2 boys and 2 girls $=N.P(X=2)$

$$=800(0.375)$$

$$=300$$

(ii) $P(\text{at least 1 boy})=P(X\geq 1)$

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

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

$$=1-4c_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4$$

$$=0.9375$$

$\therefore$ No. of families having at least 1 boy $= N.P(X\geq 1)$

$$=800(0.9375)$$

$$=750.$$

(iii) $P(\text{atmost 2 girls}) = P(\text{exactly 0 girl, 1girl, 2 girls})$

$$=P(X=4)+P(X=3)+P(X=2)$$

$$=4c_4\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0+4c_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^1+4c_4\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2$$

= 0.6875.

$\therefore$ No. of families having atmost 2 girls $= 800(0.6875) = 550$.

(iv) $P(\text{Children of both genders}) = 1-P(\text{Children of the same gender})$

$$=1-\{P(\text{all are boys})+P(\text{all are girls})\}$$

$$=1-\{P(X=4)+P(X=0)\}$$

$$=1-\left[4c_4\left(\frac{1}{2}\right)^4+4c_0\left(\frac{1}{2}\right)^4\right]$$

$$=0.875.$$

$\therefore$ No. of families having children of both genders $= 800(0.875) = 700$.

Continuous type

Uniform (or) Rectangular distribution:

A continuous random variable X is said to have a uniform distribution over an interval (a,b) if its probability density function is given by

$$f(x)=\begin{cases}\frac{1}{b-a}, a<x<b\\ 0, otherwise\end{cases}$$

Problem.1 If X is uniformly distributed with Mean1and Variance $\frac{4}{3}$, find $P[X>0]$

Solution:

If X is uniformly distributed over (a,b), then

$$E(X)=\frac{b+a}{2} \text{ and } V(X)=\frac{(b-a)^2}{12}$$

$$\therefore \frac{b+a}{2}=1\Rightarrow a+b=2$$

$$\Rightarrow \frac{(b-a)^2}{12}=\frac{4}{3}\Rightarrow (b-a)^2=16$$

$\Rightarrow a+b=2$ & $b-a=4$ We get $b=3, a=-1$

$\therefore a=-1 \,\&\, b=3$ and probability density function of x is

$$f(x)=\begin{cases}\frac{1}{4};-1<x<3\\ 0\,;Otherwise\end{cases}$$

$$P[x<0]=\int_{-1}^{0}\frac{1}{4}dx=\frac{1}{4}[x]_{-1}^{0}=\frac{1}{4}.$$

Normal distribution:

A random variable X is said to have a Normal distribution with parameters μ (mean) and σ^2 (variance) if its probability density function is given by the probability law

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

Notation: $X\sim N\left(\mu,\sigma^2\right)$ read as X is following normal distribution with mean μ and variance σ^2 are called parameter.

Problem.1 Prove that "For standard normal distribution $N(0,1)$, $M_X(t)=e^{\frac{t^2}{2}}$.

Solution:

Moment generating function of Normal distribution

$=M_X(t)=E\left[e^{tx}\right]$

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

Put $z=\frac{x-\mu}{\sigma}$ then $\sigma dz=dx,\ -\infty<Z<\infty$

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

$$=\frac{e^{\mu t}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\left(\frac{z^2}{2}-t\sigma z\right)}dz$$

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

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

$\because$ the total area under normal curve is unity, we have $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(z-t\sigma)^2}dz=1$

Hence $M_X(t)=e^{\mu t+\frac{\sigma^2t^2}{2}}$ $\therefore$ For standard normal variable $N(0,1)$

$$M_X(t)=e^{\frac{t^2}{2}}$$

Problem.2 State and prove the additive property of normal distribution.

Solution:

Statement:

If $X_1,X_2,\ldots,X_n$ are n independent normal random variates with mean $\left(\mu_1,\sigma_1{}^2\right)$, $\left(\mu_2,\sigma_2{}^2\right),\ldots\left(\mu_n,\sigma_n{}^2\right)$ then $X_1+X_2+\ldots+X_n$ also a normal random variable with mean

$\left(\sum_{i=1}^{n}\mu_i,\sum_{i=1}^{n}\sigma_i^2\right)$.

Proof:

We know that. $M_{X_1+X_2+...+X_n}(t)=M_{X_1}(t)M_{X_2}(t)...M_{X_n}(t)$

But $M_{X_i}(t)=e^{\mu_i t+\frac{t^2\sigma_i^2}{2}}$, $i=1,2....n$

$$M_{X_1+X_2+...+X_n}(t)=e^{\mu_1 t+\frac{t^2\sigma_1^2}{2}}e^{\mu_2 t+\frac{t^2\sigma_2^2}{2}}...e^{\mu_n t+\frac{t^2\sigma_n^2}{2}}$$

$$=e^{(\mu_1+\mu_2+...+\mu_n)t+\frac{(\sigma_1^2+\sigma_2^2+...+\sigma_n^2)t^2}{2}}$$

$$=e^{\sum_{i=1}^{n}\mu_i t+\frac{\sum_{i=1}^{n}\sigma_i^2 t^2}{2}}$$

By uniqueness MGF, $X_1+X_2+...+X_n$ follows normal random variable with parameter$\left(\sum_{i=1}^{n}\mu_i,\sum_{i=1}^{n}\sigma_i^2\right)$.

This proves the property.

Problem.3 X is a normal variate with $mean=30$ and $S.D=5$ Find the following $P[26\le X\le 40]$

Solution:

$X\sim N(30,5^2)$

$\therefore \mu=30\ \&\ \sigma=5$

Let $Z=\frac{X-\mu}{\sigma}$ be the standard normal variate

$$P[26\le X\le 40]=P\left[\frac{26-30}{5}\le Z\le\frac{40-30}{5}\right]$$

$$=P[-0.8\le Z\le 2]=P[-0.8\le Z\le 0]+P[0\le Z\le 2]$$

$$=P[0\le Z\ 0.8]+[0\le z\le 2]$$

$=0.2881+0.4772=0.7653$.

Problem.4 The average percentage of marks of candidates in an examination is 45 will a standard deviation of 10 the minimum for a pass is 50%.If 1000 candidates appear for the examination, how many can be expected marks. If it is required, that double that number should pass, what should be the average percentage of marks?

Solution:

Let X be marks of the candidates

Then $X\sim N(42,10^2)$

Let $z=\frac{X-42}{10}$

$$P[X>50]=P[Z>0.8]$$

$$=0.5-P[0<z<0.8]$$

$$=0.5-0.2881=0.2119$$

Since 1000 students write the test, nearly 212 students would pass the examination.

If double that number should pass, then the no of passes should be 424.

We have to find z_1, such that $P[Z>z_1]=0.424$

$\therefore P[0<z<z_1]=0.5-0.424=0.076$

From tables, $z=0.19$

$\therefore z_1=\frac{50-x_1}{10}\Rightarrow x_1=50-10z_1$

$$=50-1.9=48.1$$

The average mark should be 48 nearly.

Problem.5 Given that X is normally distribution with mean 10 and probability $P[X>12]=0.1587$. What is the probability that X will fall in the interval $(9,11)$.

Solution:

Given X is normally distributed with mean $\mu=10$.

Let $z=\frac{x-\mu}{\sigma}$ be the standard normal variate.

For $X=12, z=\frac{12-10}{\sigma}\Rightarrow z=\frac{2}{\sigma}$

Put $z_1=\frac{2}{\sigma}$

Then $P[X>12]=0.1587$

$$P[Z>Z_1]=0.1587$$

$\therefore 0.5-p[0<z<z_1]=0.1587$

$\Rightarrow P[0<z<z_1]=0.3413$

From area table $P[0<z<1]=0.3413$

$\therefore Z_1=1\Rightarrow\frac{2}{\sigma}=1$

To find $P[9<x<11]$

For $X=9,\ z=-\frac{1}{2}$ and $X=11,\ z=\frac{1}{2}$

$\therefore P[9<X<11]=P[-0.5<z<0.5]$

$$=2P[0<z<0.5]$$

$$=2\times 0.1915=0.3830$$

Problem.6 In a normal distribution 31% of the items are under 45 and 8% are over 64.Find the mean and standard deviation of the distribution.

Solution:

Let μ be the mean and σ be the standard deviation.

Then $P[X\le 45]=0.31$ and $P[X\ge 64]=0.08$

When $X=45,\ Z=\frac{45-\mu}{\sigma}=-z_1$

$\therefore z_1$ is the value of z corresponding to the area $\int_0^{z_1}\phi(z)dz=0.19$

$\therefore z_1=0.495$

$45-\mu=-0.495\sigma$ ---(1)

When $X=64, Z=\frac{64-\mu}{\sigma}=z_2$

$\therefore z_2$ is the value of z corresponding to the area $\int_0^{z_2}\phi(z)dz=0.42$

$\therefore z_2=1.405$

$64-\mu=1.405\sigma$ ---(2)

Solving (1) & (2) We get $\mu=10$ (approx) & $\sigma=50$ (approx)

UNIT –II : SAMPLING DISTRIBUTION & ESTIMATION

Population:

The group of individuals under study is called population. The population may be finite or infinite.

Sample and Sample Size:

A finite subset of statistical individuals in a population is called **Sample**. The number of individuals in a sample is called **Sample Size(n)**.

Parameter and Statistic:

A numerical measure of a population is called a population parameter or simply a parameter.

A numerical measure of the sample is called a sample statistic or simply a statistic.

Sampling distribution:

The sampling distribution of a statistic is the probability distribution of all possible values the statistic may take, when computed from random samples of same size, drawn from a specified population. Like any other distribution, a sampling distribution will have its mean, standard deviation and moments of higher order.

Standard Error:

The standard deviation of the sampling distribution of a statistic is known as itsstandard error.

Uses of Standard Error:

The magnitude of the standard error gives an index of the reliability of the estimate of the parameter. The greater the standard error of the estimate, lesser will be the reliability of the sample.

Standard error is useful for determining the probable limits or confidence limits for an unknown parameter with a specified confidence co-efficient.Standard error is also used for testing of hypothesis.

Type I error and Type II error:

Type I error: If we reject a hypothesis when it should be accepted, we say that type I error has been made.

Type II error: If we accept a hypothesis when it should be rejected, we say that a type II error has been made.

Critical region:

A region corresponding to a test statistic in the sample space which tends to rejection of H_0(Null Hypothesis) is called critical region or region of rejection.

The region complementary to the critical region is called the region of acceptance.

Level of significance:

The probability 'α' (the probability of making type I error) that a random value of the test statistic belongs to the critical region is known as the level of significance. In other words, level of significance is the size of the type I error.

The levels of significance usually employed in testing of hypothesis are 5% and 1%.

Critical values or significant values:

The value of test statistic which divides the critical (or rejection) region and acceptance region is called the critical value or significant value. It depends on the level of significance used and the alternative hypothesis.

Different types of sampling:

Non probability Samples: Judgment sample, Quota sample, Chunk sample.

Probability samples: Simple random sample, stratified sample, systematic sample, Cluster sample.

One tailed test:

When the hypothesis about the population parameter is rejected only for the value of sample statistic falling into one of the tails of the sampling distribution, then it is known as one-tailed test.

If it is right tail then it is called right-tailed test or one-sided alternative to the right and if it is on the left tail, then it is one-sided alternative to the left and called left-tailed test.

Two tailed test:

Two tailed test is one where the hypothesis about the population parameter is rejected for the value of sample statistic falling into the either tails of the sampling distribution.

Systematic sampling:

In a systematic sample, the N items in the population are partitioned into k groups by dividing the size of the population by the desired sample size n.

Stratified Sampling:

In a stratified sample, then N items in the population are first subdivided into separate subpopulations, or strata, according to some common characteristic.

Cluster Sampling:

In a cluster sample, the N items in the population are divided into several clusters so that each cluster is representative of the entire population.

Sampling Error:

Sampling errors have their origin in sampling and arise due to the fact that only a part of the population has been used to estimate populations parameters and draw inferences about them.

Estimator:

An estimator of a population parameter is a sample statistic used to estimate the parameter. An estimate of the parameter is a particular numerical value of the estimator obtained by sampling.

Different types of estimation:

There are two types of estimation. They are Point estimation and Interval estimation.

Point estimation:

When a single value is used as an estimate, the estimate is called a point estimate of the population parameter. For example, the sample mean is the sample statistic used as an estimate of population mean μ.

Interval estimation:

An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an interval estimate of the parameter.

The interval estimate or a confidence interval consists of an upper confidence limit and lower confidence limit and we assign a probability that this interval contains the unknown population parameter.

Characteristics of a good estimator:

The important properties of good statistical estimators are (i) unbiasedness (ii) efficiency (iii) consistency (iv) sufficiency.

Unbiased estimator:

An estimator is said to be unbiased if its expected value is equal to the population parameter it estimates.

Consistent estimator:

An estimator is said to be consistent if its probability of being close to the parameter it estimates increases as the sample size increases.

Efficient estimator:

An estimator is efficient if it has a relatively smaller variance.

Sufficient estimator:

An estimator is said to be sufficient if it contains all the information in the data about the parameter it estimates.

<u>FORMULAS:</u>

1. Write the confidence interval for the population mean for large samples when σ is known.

The confidence interval for μ when σ is known and sampling is done from a normal population or with a large sample, is $\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Here $\bar{x}$ - sample mean, σ - standard deviation, n – size of the sample

2. Write the confidence interval for the population mean for small samples when σ is unknown.

The confidence interval for μ when σ is not known is $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}, where\, S^2 = \frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2$

Here $\bar{x}$ - sample mean, s - standard deviation, n – size of the sample

3. Write the confidence interval for the difference between two population means for large samples when σ is known.

The confidence limits for the difference between two population means are given by

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

4. Write the confidence interval for the difference between two population means for small samples when σ is unknown.

The confidence limits for the difference between two population means are given by

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{\alpha/2} S\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ where\ S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum\left(x_i - \bar{x}\right)^2 + \sum\left(y_i - \bar{y}\right)^2\right]$$

5. Write the confidence interval for the population proportion for large samples.

The confidence interval for the population proportion P is $p \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$, where q = 1- p.

6. Write the confidence interval for the difference between two population proportions for large samples.

Confidence limits for the difference between two population proportions are

$$\left(p_1 - p_2\right) \pm Z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

7. Write the confidence interval for a mean when a finite population N is known?

$$\mu = \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

8. What is the sample size for estimating a population mean when the sample standard deviation and standard error is known?

$$n = \left(\frac{Z_\alpha . \sigma}{E}\right)^2, E - \text{standard error}$$

9. What is the sample size for estimating a population proportion when the sample standarddeviation and standard error is known?

$$n = \frac{Z_\alpha{}^2 . pq}{E^2}, E - \text{standard error}$$

<u>PROBLEMS:</u>

1. A machine produces components, which have a standard deviation of 1.6cm in length. A random sample of 64 parts is selected from the output and this sample has a mean length of 90cm. The customer will reject the part if it is either less than 88cm or more than 92cm. Does the 95% confidence interval for the true mean length of all the components produced ensure acceptance by the customer?

Solution:

$$\text{Formula for confidence interval is } \overline{\mathrm{x}} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{\mathrm{x}} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$where\,\sigma = 1.6,\ z_{\frac{\alpha}{2}} = 1.96,\ \overline{\mathrm{x}} = 90 \ and \ n = 64.$$

$$\therefore 89.61 \le \mu \le 90.39.$$

This implies that the probability that the true value of the population mean length of the components will faill in this interval is 95%.

2. A server channel monitored for an hour was found to have an estimated mean of 20 transactions transmitted per minute. The variance is known to be 4. Find the standard error. Establish an interval estimate that includes a population mean 95% of the time and 99% of the time.

Solution:

(i) Standard error $= \frac{\sigma}{\sqrt{\mathrm{n}}} = 0.2582$

(ii) $Z_{\frac{\alpha}{2}} = 1.96$`. 95% confidence interval is $\mu = \overline{\mathrm{x}} \pm Z_{\frac{\alpha}{2}} \sigma_{\bar{x}} = (19.4939, 20.5061)$

(iii) $Z_{\frac{\alpha}{2}} = 2.58$`. 99% confidence interval is $\mu = \overline{\mathrm{x}} \pm Z_{\frac{\alpha}{2}} \sigma_{\bar{x}} = (19.33, 20.67)$

3. A management consulting agency needs to estimate the average number of years of experience of executives in a given branch of management. A random sample of 28 executives gives sample mean as 6.7 years and standard deviation as 2.4 years. Give a 99% confidence interval for the average number of years of experience for all executives in this branch.

Solution:

Given n = 28, $\overline{\mathrm{x}} = 6.7 \, and \, s = 2.4.$

$t_{\frac{\alpha}{2}} = 2.861.$`. 99% confidence interval is $\mu = \overline{\mathrm{x}} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = (5.402, 7.998)$

4. In order to compare the intelligent quotient of students, two schools were selected. A random sample of 90 students was selected from each school. At school A, the mean I.Q. is 109 and standard deviation is 11. At school B, the mean I.Q. is 98 and standard deviation is 9. Construct 95% confidence interval for the difference between mean I.Q. of two schools.

Solution:

Given $\mathrm{n}_1 = n_2 = 90, \overline{\mathrm{x}_1} = 109, S_1 = 11, \overline{\mathrm{x}_2} = 98, S_2 = 9$

$z_{\frac{\alpha}{2}} = 1.96$`. 95% confidence interval is $\left(\overline{\mathrm{x}_1} - \overline{\mathrm{x}_2}\right) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = (8.06, 13.94)$

5. A sample poll of 100 voters chosen at random from all voters in a given district indicated that 55% of them were in favour of a particular candidate. Find (i) 95% and (ii) 99% confidence limits for the proportion of all the voters in favour of this candidate.

Solution: $We\,have\,n = 100.$

$Sample\ proportion\ p = 0.55\,and\ q = 0.45 \quad Z_{\alpha/2} = 1.96$

(a) $Z_{\alpha/2} = 1.96$: 95% Confidence interval for p $=$ $\mathrm{p} \pm Z_{\alpha/2}\sqrt{\dfrac{pq}{n}} = (0.4526, 0.6474)$

(b) $Z_{\alpha/2} = 2.58$: 99% Confidence interval for p $=$ $\mathrm{p} \pm Z_{\alpha/2}\sqrt{\dfrac{pq}{n}} = (0.4218, 0.6782)$

6. Two operators perform the same operation of applying a plastic coating to a part. A random sample of 100 parts from the first operator shows that 6 are non-conforming. A random sample of 200 parts from the second operator shows that 8 are non-conforming. Find a 90% confidence interval for the difference in the proportion of non-conforming parts produced by the two operators.

Solution:

$Given\,n_1 = 100, n_2 = 200, p_1 = 0.06, q_1 = 0.94, p_2 = 0.04, q_2 = 0.96,\ Z_{\alpha/2} = 1.645.$

90% confidence interval for the difference in the proportion of non - conforming parts produced is

$$(\mathrm{p}_1 - p_2) \pm Z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = (0.0275, 0.0652)$$

7. The operations manager for a large newspaper wants to determine the proportion of newspapers printed that have a non conforming attribute, such as excessive rub off, missing pages, and duplicate pages. The operations manager determines that a random sample of 200 newspapers should be selected for analysis. Suppose that, of this sample of 200, thirty five contain same type of non conformance. If the operations manager wants to have 90% confidence in estimating the true population proportion, set up the confidence interval estimate.

Solution:

$We\,have\,x = 35\,and\ n = 200.$

$Sample\ proportion\ p = \dfrac{x}{n} = 0.175, q = 0.825\ and\ \ Z_{\alpha/2} = 1.645$

90% confidence interval for $\mathrm{p} = \mathrm{p} \pm Z_{\alpha/2}\sqrt{\dfrac{pq}{n}} = (0.1308, 0.2192)$

8. The following are the average weekly losses of worker-hours due to accidents in 10 industrial plants before and after a certain safety program were put into operation:

Before	**45**	**73**	**46**	**124**	**33**	**57**	**83**	**34**	**26**	**17**
After	**36**	**60**	**44**	**119**	**35**	**51**	**77**	**29**	**24**	**11**

Find a 90% confidence interval for the mean improvement in lost worker hours.

Null Hypothesis H_0: There is no improvement between before and after the safety program.

Alternative Hypothesis H_1: There is an improvement in the performance between before and after the safety program.

From the given data

	Group 1	Group 2
Mean	53.8	48.6
Variance	1027.7333	962.9333

Stand. Dev.	32.0583	31.0312
n	10	10
t	4.0333	
degrees of freedom	9	
critical value	4.297	

Since the calculated *t* value is smaller than critical value (4.0333<4.297), we reject H_0. So the means are not significantly different.

9. In a test given to two groups of students the marks obtained were as follows

I group	18	20	36	50	49	36	34	49	61
II group	**29**	**28**	**26**	**35**	**30**	**44**	**46**		

Construct a 95% confidence interval on the mean marks secured by students of the above two groups.

Solution:

$$\text{Given } \mathrm{n}_1 = 9, n_2 = 7, \overline{\mathrm{x}_1} = 37, \overline{\mathrm{x}_2} = 34$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_i (x_i - x_1)^2 + \sum_j (x_i - x_2)^2\right] = 108.57. \Rightarrow S = 10.42.$$

$$t_{\frac{\alpha}{2}} = 1.76\text{`}. \ 95\% \text{ confidence interval is } \left(\overline{\mathrm{x}_1} - \overline{\mathrm{x}_2}\right) \pm t_{\frac{\alpha}{2}} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (6.24,12.24)$$

10. The diameter of component produced on a semi-automatic machine is known to be distributed normally with a mean of 10mm and a standard deviation of 0.1mm. If we pick up a random sample of size 5, what is the probability that the sample mean will be between 9.95mm and 10.05mm?

Solution:

$$P\left(9.95 < \overline{x} < 10.05\right) = P(-1.12 < z < 1.12) = 2P(0 < z < 1.12) = 0.7372$$

11. For a particular brand of T.V. picture tube, it is known that the mean operating life of the tubes is 1000 hours with a standard deviation of 250 hours. What is the probability that the mean for a random sample of size 25 will be between 950 and 1050 hours?

Solution:

$$P\left(950 < \overline{x} < 1050\right) = P(-1 < z < 1) = 2P(0 < z < 1) = 0.6826$$

12. Strength of wire were produced by company A has a mean of 4500kg and a standard deviation of 200kg, company B has a mean of 4000kg and a S.D. of 300kg. If 50 wires of company A and 100 wires of company B are selected at random and tested for strength. What is the probability that the sample mean strength of A will be atleast 600kg more than that of B.

Solution:

$$P\left(\overline{x} > 600\right) = P(z > 2.425) = 0.5 - 0.4925 = 0.0075$$

13. Car stereo manufacturer of A have mean life time of 1400 hours with a S.D. of 200 hours while those of manufacturer B have mean lifetime of 1200 hours with a S.D. of 100 hours. If a random

sample of 120 stereos of each manufacturer are tested. (i) What is the probability that the manufacturer of A's stereo's will have a mean life time of atleast 160 hours more than the manufacturer B's stereos (ii) and 250 hours more than the manufacturer B stereos.

Solution:

(i) $P(\bar{x} > 160) = P(z > -1.95) = 0.5 + 0.4750 = 0.9750$

(ii) $P(\bar{x} > 250) = P(z > 2.45) = 0.5 - 0.4929 = 0.0071$

14. If two proportions 10% of machine produced by a company A are defective and 5% of machine produced by a company B are defective. A random sample of 250 machines is taken from company A and has the random sample of 300 machines from company B. What is the probability that the difference in sample proportion is less than or equal to 0.02.

Solution:

$$P(p_1 - p_2 < 0.02) = P(z < -1.32) = 0.5 - P(0 < z < 1.32) = 0.5 - 0.4066 = 0.0934$$

15. A random sample of 500 toys was taken from a consignment and 65 were found to be defective. Find the percentage of defective toys in the consignment.

Solution:

$n = 500;\ X = 65.\ p = 0.13$

The limits for the population proportion P are given by $p \pm 1.96\sqrt{\frac{pq}{n}} = (0.159, 0.101)$

The percentage of defective toys in the consignment lies betwwen and 10.1% and 15.9%.

16. What are the different types of sampling methods? Also write short notes on different types of sampling?

Solution:

Simple random sampling, Stratified sampling, Cluster sampling, Judgment sampling and Quota sampling.

Technique	**Descriptions**	**Advantages**	**Disadvantages**
Simple random	Random sample from whole population	Highly representative if all subjects participate; the ideal	Not possible without complete list of population members; potentially uneconomical to achieve; can be disruptive to isolate members from a group; time-scale may be too long, data/sample could change
Stratified random	Random sample from identifiable groups (strata), subgroups, etc.	Can ensure that specific groups are represented, even proportionally, in the sample(s) (e.g., by gender), by selecting individuals from strata list	More complex, requires greater effort than simple random; strata must be carefully defined
Cluster	Random samples of successive clusters of subjects (e.g., by institution) until small groups are chosen as units	Possible to select randomly when no single list of population members exists, but local lists do; data collected on groups may avoid introduction of confounding by isolating members	Clusters in a level must be equivalent and some natural ones are not for essential characteristics (e.g., geographic: numbers equal, but unemployment rates differ)

Stage	Combination of cluster (randomly selecting clusters) and random or stratified random sampling of individuals	Can make up probability sample by random at stages and within groups; possible to select random sample when population lists are very localized	Complex, combines limitations of cluster and stratified random sampling
Purposive	Hand-pick subjects on the basis of specific characteristics	Ensures balance of group sizes when multiple groups are to be selected	Samples are not easily defensible as being representative of populations due to potential subjectivity of researcher
Quota	Select individuals as they come to fill a quota by characteristics proportional to populations	Ensures selection of adequate numbers of subjects with appropriate characteristics	Not possible to prove that the sample is representative of designated population
Snowball	Subjects with desired traits or characteristics give names of further appropriate subjects	Possible to include members of groups where no lists or identifiable clusters even exist (e.g., drug abusers, criminals)	No way of knowing whether the sample is representative of the population
Volunteer, accidental, convenience	Either asking for volunteers, or the consequence of not all those selected finally participating, or a set of subjects who just happen to be available	Inexpensive way of ensuring sufficient numbers of a study	Can be highly unrepresentative

17. Given a sample mean of 83, a sample standard deviation of 12.5 and a sample size of 22, test the hypothesis that the value of the population mean is 70 against the alternative that it is more than 70. Use the 0.025 level of significance.

Here the sample size, n = 22 < 30. Hence the sample is small sample. Given $\bar{x} = 83$, $\mu = 70$, $s = 12.5$

Null Hypothesis H_0: There is no significant different between sample mean and population mean.

Alternative Hypothesis H_1: $\mu > 70$.

Degrees of freedom: n–1 = 21.

The test statistic is, $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n-1}}$

$$= \frac{83 - 70}{12.5/\sqrt{21}} = 4.7659$$

Tabulated value of t at 5% level with 21 degrees of freedom for single tailed test is 1.72. Here calculated value > tabulated value, we reject H_0.

Therefore $\mu > 70$.

18. All the 0.10 level of significance, can we conclude that the following 400 observations follow a Poisson distribution with $\lambda = 3$?

No of arrivals per hr	**0**	**1**	**2**	**3**	**4**	**5 or more**
No of hrs	**20**	**57**	**98**	**85**	**78**	**62**

Given λ = 3, N = 400. The expected frequency is $P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}*N$

Null Hypothesis H_0: The given data fit Poisson distribution
Alternative Hypothesis H1: The given data does not fit Poisson distribution

X	O	E	$(O\text{-}E)^2$	$(O\text{-}E)^2$ / E
0	20	19.912	0.0077	0.0004
1	57	59.736	7.4857	0.1313
2	98	89.604	70.4928	0.7193
3	85	89.604	21.1968	0.2494
4	78	67.203	116.5752	1.4946
5	62	40.3218	469.9444	7.5797
	Total			10.17471

Degrees of freedom , n-2 = 6-2 = 4
Tabulated value = 7.779
Since the calculated value is greater than the tabulated value, we reject H_0.
Therefore the given data does not fit Poisson distribution.

19. If $X_1, X_2, ..., X_n$ are Poisson variates with parameter $\lambda = 2$, use the central limit theorem to estimate $P(120 \le s_n \le 160)$ where s_n = $X_1 + X_2 + ... + X_n$ and n=75

Sol:

Given $\mu = 2, \sigma^2 = 2$ $(For\ \ poisson\ distribution,\ mean = \text{var}\,iance = \lambda)$

Now, $n\mu = 75 \times 2 = 150,\ \&\ \ n\sigma^2 = 75 \times 2 = 150 \Rightarrow \sigma\sqrt{n} = \sqrt{150}$

By Central limit theorem, $S_n : N(n\mu, \sigma\sqrt{n}) \approx N(150, \sqrt{150})$

To find P(120< S_n <160):

Let $z = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - 150}{\sqrt{150}}$, (since z is standard normal variate)

If S_n =120, then $z = \frac{120-150}{\sqrt{150}} = -2.45$ and if S_n =160, $z = \frac{160-150}{\sqrt{150}} = 0.85$,

$\therefore$ P(120< S_n <160)=P(-2.45<z<0.85)
=P(-2.45<z<0)+P(0<z<0.85)
=P(0<z<2.45)+ P(0<z<0.85)
=0.4927+0.2939
=0.7866

Unit-III Testing of Hypothesis

Population:

A population in statistics means a set of object. The population is finite or infinite according to the number of elements of the set is finites or infinite.

Sampling:

A sample is a finite subset of the population. The number of elements in the sample is called size of the sample.

Large and small sample:

The number of elements in a sample is greater than or equal to 30 then the sample is called a large sample and if it is less than 30, then the sample is called a small sample.

Parameters:

Statistical constant like mean μ, variance σ^2 , etc., computed from a population are called parameters of the population.

Statistics:

Statistical constants like $\overline{x}$, variance S^2, etc., computed from a sample are called samlple staticts or statistics.

POPULATION (PARAMETER)	**SAMPLE (STATISTICS)**
Population size=N	Sample size=n
Population mean= μ	Sample mean= $\overline{x}$
Population s.d.=σ	Sample s.d.=S
Population proportion= P	Sample proportion= p

<u>Tests of significance or Hypothesis Testing:</u>

Statistical Hypothesis:

In making statistical decision, we make assumption, which may be true or false are called Statistical Hypothesis.

Null Hypothesis(H_0):

For applying the test of significance, we first setup a hypothesis which is a statement about the population parameter. This statement is usually a hypothesis of no true difference between sample statistics and population parameter under consideration and so it is called null hypothesis and is denoted by H_0.

Alternative Hypothesis (H_1):

Suppose the null hypothesis is false, then something else must be true. This is called an alternative hypothesis and is denoted by H_1.

Eg. If H_0 is population mean μ=300, then H_1 is $\mu \neq 300$ $(ie. \mu < 300 \; or \; \mu > 300)$ or $H_1 \; is \, \mu > 300 \; or \; H_1 \; is \, \mu < 300$. So any of these may be taken as alternative hypothesis.

<u>Error in sampling:</u>

After applying a test of significance a decision is to be taken to accept or reject the null hypothesis H_0.

Type I error: The rejection of the null hypothesis H_0 when it is true is called type I error.

Type II error: The acceptance of the null hypothesis H_0 when it is false is called type II error.

Level of significance:

The probability of type I error is called level of significance of the test and it is denoted by α. We usually take either α=5% or α=1%.

One tailed and Two tailed test:

If θ_0 is a population parameter and θ is the corresponding sample statistics and if we setup the null hypothesis $H_0 : \theta = \theta_0$, then the alternative hypothesis which is complementary to H_0 can be anyone of the following:

(i) $H_1 : \theta \neq \theta_0 \; (\theta < \theta_0 \;\; or \;\; \theta > \theta_0)$ (ii) $H_1 : \theta < \theta_0$ (iii) $H_1 : \theta > \theta_0$

Alternative hypotheis, whereas H_1 given in (ii) is called a left-tailed test. And (iii) is called a right tailed test.

Level of significance:

The probability of Type I error is called the level of significance of the test and is denoted by α .

Critical region:

For a test statistic, the area under the probability curve, which is normal is divided into two region namely the region of acceptance of H_0 and the region of rejection of H_0. The region in which H_0 is rejected is called critical region. The region in which H_0 is accepted is called acceptance region.

Procedure of Testing of Hypothesis:

(i) State the null hypothesis H_0

(ii) Decide the alternative hypothesis H_1 (ie, one tailed or two tailed)

(iii) Choose the level of significance α (α=5% or α=1%).

(iv) Determine a suitable test statistic.

$$\text{Test statistic } = \frac{t - E(t)}{S.E \;\; of \;\; (t)}$$

(v) Compute the computed value of $|z|$ with the table value of z and decide the acceptane or the rejection of H_0.

If $|z|$<1.96, H_0 may be accepted at 5% level of significance. If $|z|$>1.96, H_0 may be rejection at 5% level of significance.

If $|z|$<2.58, H_0 may be accepted at 1% level of significance. If $|z|$>2.58, H_0 may be rejection at 1% level of significance.

For a single tail test(right tail or left tail) we compare the computed value of $|z|$ with 1.645(at 5% level) and 2.33(at 1% level) and accept or reject H_0 accordingly.

Test of significance of small sample:

When the size of the sample (n) is less than 30, then that sample is called a small sample. The following are some important tests for small sample,

(I) students t test

(II) F-test

(III) χ^2 -test

I **Student t test**

(i). Test of significance of the difference between sample mean and population mean

(ii). Test of significance of the difference between means of two small samples

(i) **Test of significance of the difference between sample mean and population mean:**

The studemts 't' is defined by the statistic $t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$ where $\bar{x}$ =sample mean, μ=population mean, S=standard deviation of sample,
n= sample size.

Note:

If standard deviation of sample is not given directly then, the static is given by $t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$, where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$

Confident Interval:

The confident interval for the population mean for small sample is $\bar{x} \mp t_\alpha \frac{s}{\sqrt{n}}$

$$\Rightarrow \left(\bar{x} - t_\alpha \frac{s}{\sqrt{n}}, \bar{x} + t_\alpha \frac{s}{\sqrt{n}}\right)$$

Working Rule:

(i) Let $H_0 : \mu = \bar{x}$ (there is no significant difference between sample mean and population mean)

$H_1 : \mu \neq \bar{x}$ (there is no significant difference between sample mean and population mean)(Two tailed test)

Find $t = \frac{\bar{x} - \mu}{S/\sqrt{n-1}}$.

Let t_α be the table value of t with v=n-1 degrees of freedom at α % level of significance.

Conclusion:

If $|t| < t_\alpha$, H_0 is accepted at α % level of significance.

If $|t| > t_\alpha$, H_0 is rejected at α % level of significance.

Problem:

1. **The mean lifetime of a sample of 25 bulbs is found as 1550h, with standard deviation of 120h. The company manufacturing the bulbs claims that the average life of their bulbs is 1600h. Is the claim acceptable at 5% level of significance?**

Solution:

Given sample size n=25, mean $\bar{x}$=1550, S.D.(S)=120, population mean μ=1600

Let $H_0 : \mu = 1600$ (the claim is acceptable)

$H_1 : \mu \neq 1600 \ (\mu \neq \bar{x})$(two tailed test)

Under H_0 , the test statistic is $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}} = \dfrac{1550 - 1600}{120/\sqrt{25}} = -2.0833$

$\therefore |t| = 2.0833$

From the table, for v=24, $t_{0.05}$=2.064. Since $|t| > t_{0.05}$

$\therefore H_0$ is rejected

<u>Conclusion:</u> The claim is not acceptable.

2. **Test made on the breaking strength of 10 pieces of a metal gave the following results: 578,572,570,568,572,570,570,572,596, and 584kg. Test if the mean breaking strength of the wire can be assumed as 577kg.**

Solution:

let us first compute sample mean $\bar{x}$ and sample standard deviation S and then test if $\bar{x}$ differs significantly from the population mean μ=577.

x	$x - \bar{x}$	$\left(x - \bar{x}\right)^2$
578	2.8	7.84
572	-3.2	10.24
570	-5.2	27.04
568	-7.2	51.84
572	-3.2	10.24
570	-5.2	27.04
570	-5.2	27.04
572	-3.2	10.24
596	20.8	432.64
584	8.8	77.44
5752	**0**	**681.6**

Where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5752}{10} = 575.2,$$

$$S^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1} = \frac{681.6}{9} = 75.733$$

Let $H_0 : \mu = \bar{x}$,

$H_1 : \mu \neq \bar{x}$

Under H_0 , the test statistic is $t = \frac{\bar{x}-\mu}{S/\sqrt{n}} = \frac{572.2-577}{\sqrt{75.733}/\sqrt{10}} = -1.74$

$\therefore |t| = 1.74$

Tabulated value of t for v=9 degrees of freedom $t_{0.05}$ =2.262

Since $|t| < t_{0.05}$. $\therefore H_0$ is accepted

Conclusion:

$\therefore$ The mean breaking strength of the wire can be assumed as 577kg at 5% level of significance.

3. **A random sample of 10 boys had the following I.Q's: 70, 120, 110, 101, 88, 83, 95, 98,107,100. Do these data support the assumption of a population mean I.Q of 100 ? Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.**

Solution:

Given $\mu = 100,\ n = 10$

Null Hypothesis:

$H_0 : \mu = 100$ i.e., The data are consist with the assumption of men IQ of 100 in the population

Alternate Hypothesis:

$H_1 : \mu \neq 100$ i.e., The data are consist with the assumption of men IQ of 100 in the population

Level of Significance : $\alpha = 5\% \Rightarrow \alpha = 0.05$

Test Statistic :

$$t = \frac{\bar{x}-\mu}{S/\sqrt{n}}$$

where $S^2 = \frac{1}{n-1}\sum (x-\bar{x})^2$

$$\bar{x} = \frac{\sum x}{n} = \frac{70+120+110+101+88+83+95+98+107+100}{10} = \frac{972}{10} = 97.2$$

$$S^2 = \frac{1}{10-1}\begin{bmatrix} (70-97.2)^2 + (120-97.2)^2 + (110-97.2)^2 + (101-97.2)^2 + (88-97.2)^2 \\ +(83-97.2)^2 + (95-97.2)^2 + (98-97.2)^2 + (107-97.2)^2 + (100-97.2)^2 \end{bmatrix}$$

$$S^2 = \frac{1}{9}(1833.6) = 203.73 \Rightarrow S = 14.2734$$

$$t=\frac{97.2-100}{14.2734\Big/\sqrt{10}}=\frac{2.8}{4.5136}=0.6203$$

Table value : $t_{\alpha,n-1}=t_{5\%,10-1}=t_{0.05,9}=2.262$ **(Two –tailed test)**

Conclusion :

Here $t>t_\alpha$

i.e., The table value >calculated value,

$\therefore$ we accept the null hypothesis and conclude that the data are consistent with the assumption of mean I.Q of 100 in the population.

To find the confidence limit:

$$\left(\overline{x}\mp t_\alpha\frac{S}{\sqrt{n}}\right)=\left(97.2\mp 2.262\times\frac{14.2734}{\sqrt{10}}\right)=\left(97.2\mp(2.262)(4.514)\right)=(86.99,107.41)$$

A reasonable range in which most of the mean I.Q. values of samples of 10 boys lies $(86.99,107.41)$

4. **A random sample of 16 values from a normal population showed a mean of 41.5 inches and the sum of squares of deviations from this mean equal to 135 square inches. Show that the assumption of a mean of 43.5 inches for the population is not reasonable. Obtain 95 percent and 99 percent confidence limits for the same.**

Solution:

Given $\overline{x}=41.5,\ \mu=43.5,\ n=16$

Sum of squares of deviations from mean= $\sum\left(x-\overline{x}\right)^2=135$

The parameter of interest is μ .

Null Hypothesis H$_0$: μ =43.5 i.e., the assumption of a mean of 43.5 inches for the population is reasonable.

Alternative Hypothesis H$_1$: $\mu\neq 43.5$ i.e., the assumption of a mean of 43.5 inches for the population is not reasonable.

Level of significance: (i) α =5%= 0.05, degrees of freedom = 16–1=15

(ii) α =1% =0.01, degrees of freedom = 16–1=15

Test Statistic : $t=\dfrac{\overline{x}-\mu}{\dfrac{S}{\sqrt{n}}}$

where $S^2=\dfrac{1}{n-1}\sum(x-\overline{x})^2=\dfrac{1}{16-1}135=9\ \Rightarrow S=9$

$$t=\frac{41.5-43.5}{\dfrac{3}{\sqrt{16}}}=\frac{-8}{3}=-2.667\quad\Rightarrow|\,t\,|=2.667$$

Conclusion:

(i) Since $|\,t\,|$=2.667 > 2.131 so we reject H$_0$ at 5% level of significance.

So we conclude that the assumption of mean of 43.5 inches for the population is not reasonable.

(ii) Since $|\,t\,|$=2.667 < 2.947 so we accept H$_0$ at 1% level of significance.

So we conclude that the assumption is reasonable.

95% confidence limits:

$$\left(\bar{x} \mp t_\alpha \frac{S}{\sqrt{n}}\right) = \left(41.5 \mp \left(2.947 \times \frac{3}{4}\right)\right) = \left(41.5 \mp 1.5983\right) = (39.9, 43.09)$$

$\therefore 39.902 < \mu < 43.098$

99% confidence limits:

$$\left(\bar{x} \mp t_\alpha \frac{S}{\sqrt{n}}\right) = \left(41.5 \mp \left(2.947 \times \frac{3}{4}\right)\right) = \left(41.5 \mp 2.2101\right) = (39.29, 43.71)$$

$\therefore 39.29 < \mu < 43.71$

5. **Ten oil tins are taken at random from an automatic filling machine. The mean weight of the tins is 15.8 kg and standard deviation is 0.5 kg. Does the sample mean differ significantly from the intended weight of 16 kg?**

Solution:

Given $\bar{x} = 15.8,\ \mu = 16,\ s = 0.50,\ n = 10$

Null Hypothesis H_0: $\mu = 16$ the sample mean weight is not different from the intended weight.

Alternative Hypothesis H_1: $\mu \neq 16$ i.e., the sample mean weight is not different from the intended weight.

Level of significance: α =5%= 0.05, degrees of freedom = 10-1=9

Test Statistic : $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$

$$t = \frac{15.8 - 16}{\dfrac{0.50}{\sqrt{10}}} = \frac{-0.2}{0.1581} = -1.27 \quad \Rightarrow |\,t\,| = 1.27$$

Critical value : The critical value of *t* at 5% level of significance with degrees of freedom 9 is 2.26

Conclusion:

Here calculated value < table value.

so we accept H_0 at 5% level of significance.

Hence the sample mean weight is not different from the intended weight.

(ii) **Test of significance of the difference between means of two small samples:**

To test the significance of the difference between the means $\bar{x}_1$ and $\bar{x}_2$ of sample of size n_1 and n_2.

Under H_0 , the test statistic is $t = \dfrac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$,

$$where\ S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}\ \ or\ S^2 = \frac{\sum\left(x_1 - \bar{x}_1\right)^2 + \sum\left(x_2 - \bar{x}_2\right)^2}{n_1 + n_2 - 2}\ (if\ s_1, s_2\ is\ not\ given\ directly)$$

Degrees of freedom(df) v = $n_1 + n_2$ -2

Note:

If $n_1 = n_2$ =n and if the pairs of values x_1 and x_2 are associated in some way (or correlated).

Then we use the statistic is $t = \frac{\overline{d}}{S/\sqrt{n-1}}$, *where* $\overline{d} = \frac{\sum d}{n}$ *and* $S^2 = \frac{\sum\left(d-\overline{d}\right)^2}{n}$

Degrees of freedom v = n-1

Confident Interval:

The confident interval for difference between two population means for small sample is $\left(\overline{x}_1 - \overline{x}_2\right) \mp t_\alpha \; S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

Problem:

1. **Samples of two types of electric bulbs were tested for length of life and the following data were obtainded.**

Sample	**Size**	**Mean**	**S.D**
I	**8**	**1234h**	**36h**
II	**7**	**1036h**	**40h**

Is the difference in the means sufficient to warrant that type I bulbs are superior type II bulbs?

Solution:

Here $\overline{x_1}$ =1234, $\overline{x_2}$ =1036, n_1 =8, n_2 =7, s_1 =36, $s_2 = 40$

Let $H_0 : \overline{x_1} = \overline{x_2}$,

$H_1 : \overline{x_1} > \overline{x_2}$ (ie. Type I bulbs are superior to type II bulbs) (one tail test)

Under H_0 , the test statistic is $t = \frac{\overline{x_1} - \overline{x_2}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$,

$$where\ S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = 40.7317$$

$$\therefore t = \frac{1234-1036}{40.7317\sqrt{\frac{1}{8}+\frac{1}{7}}} = 9.39$$

Degrees of freedom v = $n_1 + n_2$ -2=13

Tabulated value of t for 13 d.f. at 5% level of significance is $t_{0.05}$ =1.77

Since $|t| > t_{0.05}$. $\therefore H_0$ is rejected. H_1 is accepted.

Conclusion:

Type I bulbs may be regarded superior to type II bulbs at 5% level of significance.

2. Two independent sample of size 8 and 7 contained the following value:

Sample I	**19**	**17**	**15**	**21**	**16**	**18**	**16**	**14**
Sample II	**15**	**14**	**15**	**19**	**15**	**18**	**16**	

Is the difference between the sample means significant?

Solution:

x_1	$x_1 - \overline{x_1}$	$\left(x_1 - \overline{x_1}\right)^2$	x_2	$x_2 - \overline{x_2}$	$\left(x_2 - \overline{x_2}\right)^2$
19	2	4	15	-1	1
17	0	0	14	-2	4
15	-2	4	15	-1	1
21	4	16	19	3	9
16	-1	1	15	-1	1
18	1	1	18	2	4
16	-1	1	16	0	0
14	-3	9			
136	**0**	**36**	**112**	**0**	**20**

$$\overline{x_1} = \frac{\sum x_1}{n_1} = \frac{136}{8} = 17, \overline{x_2} = \frac{\sum x_2}{n_2} = \frac{112}{7} = 16$$

$$S^2 = \frac{\sum\left(x_1 - \overline{x_1}\right)^2 + \sum\left(x_2 - \overline{x_2}\right)^2}{n_1 + n_2 - 2} = \frac{36+20}{8+7-2} = 4.3076 \Rightarrow S = 2.0754$$

Let $H_0 : \overline{x_1} = \overline{x_2}$,

$H_1 : \overline{x_1} \neq \overline{x_2}$ (Two tailed test)

Under H_0 , the test statistic is $t = \dfrac{\overline{x_1} - \overline{x_2}}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{17-16}{2.0754\sqrt{\dfrac{1}{8} + \dfrac{1}{7}}} = 0.9309$

$|t| = 0.9309$

Degrees of freedom v = v = $n_1 + n_2$ -2=13

From the 't' table, v = 13 degrees freedom at 5% level of significance is $t_{0.05}$ =2.16

Since $|t| < t_{0.05} \therefore H_0$ is accepted

Conclusion:

The two sample mean do not differ significantly at 5% level of significance.

3. The following data represent the biological values of protein from cow's milk and buffalo's milk:

Cow's milk	**1.82**	**2.02**	**1.88**	**1.61**	**1.81**	**1.54**
Buffalo's milk	**2.00**	**1.83**	**1.86**	**2.03**	**2.19**	**1.88**

Examine whether the average values of protein in the two samples significantly differ at 5% level.

Solution:

Given $n_1 = n_2 = 6$

H_0: $\mu_1 = \mu_2$ There is no significant difference between the means of the two samples.

H_1: $\mu_1 \neq \mu_2$ There is a significant difference between the means of the two samples.

Test Statistic: $t = \dfrac{\bar{x} - \bar{y}}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

x	y	$x-\bar{x}$ $x-1.78$	$(x-\bar{x})^2$	$y-\bar{y}$ $y-1.965$	$(y-\bar{y})^2$
1.82	2	0.04	0.0016	0.035	0.00123
2.02	1.83	0.24	0.0576	-0.135	0.01823
1.88	1.86	0.1	0.01	-0.105	0.01102
1.61	2.03	-0.17	0.0289	0.065	0.00425
1.81	2.19	0.03	0.0009	0.225	0.0506
1.54	1.88	-0.24	0.0576	-0.085	0.00723
Total 10.68	11.79		0.1566		0.09256

$$\bar{x} = \frac{\sum x}{n_1} = \frac{10.68}{6} = 1.78\,;\ \bar{y} = \frac{\sum y}{n_2} = \frac{11.79}{6} = 1.965$$

$$S^2 = \frac{1}{6+6-2}\left[0.1566 + 0.09256\right] = (0.1)(0.2492) = 0.0249 \quad \Rightarrow S = 0.1578$$

$$t = \frac{1.78 - 1.956}{(0.1578)\sqrt{\frac{1}{6} + \frac{1}{6}}} = \frac{-0.176}{(0.1578)(0.5774)} = \frac{-0.176}{0.0911} = 1.9319$$

Critical value: The critical value of t at 5% level of significance with degrees of freedom 10 is 2.228

Here calculated value < table value, we accept H_0

(i.e,) The difference between the mean protein values of the two varieties of milk is not significant at 5% level.

4. **The following data relate to the marks obtaind by 11 students in 2 test, one held at the beginning of a year and the other at the end of the year intensive coaching.**

Test 1	**19**	**23**	**16**	**24**	**17**	**18**	**20**	**18**	**21**	**19**	**20**
Test 2	**17**	**24**	**20**	**24**	**20**	**22**	**20**	**20**	**18**	**22**	**19**

Do the data indicate that the students have benefited by coaching?

Solution:

The given data relate to the marks obtained in 2 tests by the same set of students. Hence the marks in the 2 set can be regarded as correlated.

We use t-test for paired values.

Let $H_0 : \overline{x_1} = \overline{x_2}$,

$H_1 : \overline{x_1} < \overline{x_2}$ (one tailed test)

x_1	x_2	d = x_1 - x_2	$d^2 = \left(\overline{x_1} - \overline{x_2}\right)^2$	d - $\overline{d}$	$\left(d - \overline{d}\right)^2$
19	17	2	4	3	9
23	24	-1	1	0	0
16	20	-4	16	-3	9
24	24	0	0	1	1
17	20	-3	9	-2	4
18	22	-4	16	-3	9
20	20	0	0	1	1
18	20	-2	4	-1	1
21	18	3	9	4	16
19	22	-3	9	-2	4
20	19	1	1	2	4
		-11			**58**

$$\overline{d} = \frac{\sum d}{n} = \frac{-11}{11} = 1 \quad S^2 = \frac{\sum\left(d - \overline{d}\right)^2}{n} = \frac{58}{11} = 5.272$$

the test statistic is $t = \dfrac{\overline{d}}{S/\sqrt{n-1}} = \dfrac{-1}{\sqrt{5.272}/\sqrt{10}} = -1.377 \Rightarrow |t| = 1.377$

from the table, v = n-1 = 10 (d.f.), $t_{0.05}$ =1.812

Since $|t| < t_{0.05}$ $\therefore H_0$ is accepted

Conclusion:

The students have not benefitted by coaching.

5. Ten Persons were appointed in the officer cadre in an office. Their performance was noted by giving a test and the marks were recorded out of 100.

Employee	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Before training	**80**	**76**	**92**	**60**	**70**	**56**	**74**	**56**	**70**	**56**
After training	**84**	**70**	**96**	**80**	**70**	**52**	**84**	**72**	**72**	**50**

By applying the t-test, can it be concluded that the employees have been benefited by the training?

Solution:

Null Hypothesis H₀: $\mu_1 = \mu_2$ i.e., the employees have not been benefited by the training.

Alternative Hypothesis H₁: $\mu_1 \neq \mu_2$ i.e., the employees have been benefited by the training.

Level of significance: $\alpha = 5\%$ = 0.05 (one tailed test)

Test Statistic : $t = \dfrac{\bar{d}}{\dfrac{S}{\sqrt{n}}}$

where $S^2 = \dfrac{1}{n-1}\sum(\mathrm{d}-\bar{d})^2$ & $\bar{d} = \dfrac{\sum d}{n}$

Employees	**Before**	**After**	**d**	$(d-\bar{d})^2$
A	80	84	-4	0
B	76	70	6	100
C	92	96	-4	0
D	60	80	-20	256
E	70	70	0	16
F	56	52	4	64
G	74	84	-10	36
H	56	72	-16	144
I	70	72	-2	4
J	50	50	6	100
Total			44	44.4

$$\bar{d} = \frac{\sum d}{n} = \frac{-40}{10} = -4$$

$$S^2 = \frac{1}{n-1}\sum(\mathrm{d}-\bar{d})^2 = \frac{1}{9}(720) = 80$$

$$t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}} = \frac{-4}{8.94/\sqrt{10}} = -1.414 \Rightarrow |t| = 1.414$$

Critical value : The critical value of tat 5% level of significance with degrees of freedom 9 is 1.83

Conclusion:

Here calculated value < table value.

so we accept H₀

Hence the employees have not been benefited by the training.

6. The weight gains in pounds under two systems of feeding of calves of 10 pairs of identical twins is given below.

Twin pair	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Weight gains under System A	**43**	**39**	**39**	**42**	**46**	**43**	**38**	**44**	**51**	**43**
Sytem B	**37**	**35**	**34**	**41**	**39**	**37**	**37**	**40**	**48**	**36**

Discuss whether the difference between the two systems of feeding is significant.

Solution:

Null Hypothesis H0: $\mu_1 = \mu_2$ i.e., there is no significance difference between the two system of feedings

Alternative Hypothesis H1: $\mu_1 \neq \mu_2$ i.e., there is significance difference between the two systems of feedings.

Level of significance: $\alpha = 5\%$ = 0.05 (Two tailed test)

Test Statistic : $t = \dfrac{\bar{d}}{\dfrac{S}{\sqrt{n}}}$

where $S^2 = \dfrac{1}{n-1}\sum(\mathrm{d}-\bar{d})^2$ & $\bar{d} = \dfrac{\sum d}{n}$

Twin Pair	System A x	System B y	$d = x - y$	$(d-\bar{d})^2$
1	43	37	6	2.56
2	39	35	4	0.16
3	39	34	5	0.36
4	42	41	1	11.56
5	46	39	7	6.76
6	43	37	6	2.56
7	38	37	1	11.56
8	44	40	4	0.16
9	51	48	3	1.96
10	43	36	7	6.76
Total			44	44.4

$$\bar{d} = \frac{\sum d}{n} = \frac{44}{10} = 4.4$$

$$S^2 = \frac{1}{n-1}\sum(\mathrm{d}-\bar{d})^2 = \frac{1}{9}(44.4) = 4.93 \quad \Rightarrow S = 2.08$$

$$t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}} = \frac{4.4}{2.08/\sqrt{10}} = 6.68$$

Critical value : The critical value of tat 5% level of significance with degrees of freedom 9 is 2.62

Conclusion:

Here calculated value < table value.

so we accept H_0

Hence there is no significance difference between the two systems of feedings.

II **F-test**

(i) To test whether if there is any significant difference between two estimates of population variance

(ii) To test if the two sample have come from the same population.

We use F-test:

The test statistic is given by $F = \frac{S_1^2}{S_2^2}, if\ S_1^2 > S_2^2$

Where $S_1^2 = \frac{n_1 s_1^2}{n_1 - 1}$ [n_1 is the first sample size] and $S_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$ [n_2 is the second sample size]

The degrees of freedom (v_1, v_2)=($n_1 - 1\ n_2 - 1$)

Note :

1. If $S_1^2 < S_2^2$ then $F = \frac{S_2^2}{S_1^2}$ (always F > 1)

2. To test whether two independent samples have been drawn from the same normal population, we have to test

i) Equality of population means using t-test or z-test, according to sample size.

ii) Equality of population variances using F-test

Problem:

1. A sample of size 13 gave an estimated population variance of 3.0, while another sample of size 15 gave an estimate of 2.5. Could both sample be from population with the same variance?

Solution:

Given n_1=13, n_2=15, S_1^2=3.0, $S_2^2 = 2.5$

Let $H_0 : S_1^2 = S_2^2$ (the two samples have been drawn from populations with same variance}

$H_1 : S_1^2 \neq S_2^2$

The test statistics is $F = \frac{S_1^2}{S_2^2} = \frac{3}{2.5} = 1.2$

From the table, with degrees of freedom v = ($n_1 - 1\ n_2 - 1$) = (12, 14)

$F_{0.05} = 2.53$ Since $F < F_{0.05}$ $\therefore H_0$ is accepted

Conclusion:

The two sample could have come from two normal population with the same variance.

2. **Two sample of size 9 and 8 give the sums of squares of deviations from their respective means equal to 160 and 91 respectively. Could both samples be from populations with the same variance?**

Solution:

Given n_1=9, n_2=8, $\sum\left(x-\overline{x}\right)^2 = 160$, $\sum\left(y-\overline{y}\right)^2 = 91$

$$S_1^2 = \frac{\sum\left(x-\overline{x}\right)^2}{n_1-1} = \frac{160}{8} = 20,\ S_2^2 = \frac{\sum\left(y-\overline{y}\right)^2}{n_2-1} = \frac{91}{7} = 13$$

Let $H_0 : \sigma_1^2 = \sigma_2^2$ (the two normal populations have the same variance}

$H_1 : \sigma_1^2 \neq \sigma_2^2$

The test statistics is $F = \frac{S_1^2}{S_2^2} = \frac{20}{13} = 1.538$

From the table, with degrees of freedom v = ($n_1 - 1\ n_2 - 1$) = (8,7)

$F_{0.05} = 3.73$ Since $F < F_{0.05}$ $\therefore$ H_0 is accepted

Conclusion:

The two sample could have come from two populations with the same variance.

3. **Two random samples gave the following data:**

Sample	Size	Mean	Variance
I	**8**	**9.6**	**1.2**
II	**11**	**16.5**	**2.5**

Cane we conclude that the two samples have been drawn from the same normal population?

Solution:

The two samples have been drawn from the same normal population we have to check

(i) the variance of the population do not differ significantly by F-test.

(ii) the sample means do not differ significantly by t-test.

(i) F-test:

Given n_1=8, n_2=11, s_1^2=1.2, $s_2^2 = 2.5$, $\overline{x_1}$=9.6, $\overline{x_2}$=16.5

$$S_1^2 = \frac{n_1 s_1^2}{n_1-1} = \frac{8(1.2)}{7} = 1.37 \quad S_2^2 = \frac{n_2 s_2^2}{n_2-1} = \frac{11(2.5)}{10} = 2.75$$

Let $H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

The test statistics is $F = \frac{S_2^2}{S_1^2}$ $(\sin ce\ S_1^2 < S_2^2)$

$$= \frac{2.75}{1.37} = 2.007$$

From the table, $F_{0.05}\left(n_2-1, n_1-1\right) = F_{0.05}(10,7) = 3.63$

Since $F < F_{0.05}$ $\therefore$ H_0 is accepted

(ii) t-test:(Equality of means)

Let $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Under H_0 , the test statistic is $t = \dfrac{\overline{x_1} - \overline{x_2}}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$,

$$where\ S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{8(1.2) + 11(2.5)}{8 + 11 - 2}} = 1.4772$$

$$t = \frac{9.6 - 16.5}{1.4772\sqrt{\dfrac{1}{8} + \dfrac{1}{11}}} = -10.0525 \Rightarrow |t| = 10.0525$$

From the table ,with degrees of freedom $n_1 + n_2$ -2=17, $t_{0.05}$ =2.110

$\sin ce\ |t| > t_{0.05}\ \therefore H_0$ is rejected ie. $\mu_1 \neq \mu_2$

Conclusion:

$\therefore$ The two samples could not have been drawn from the same normal population.

4. **Two independent samples of 5 and 6 items respectively had the following values of the following values of the variable:**

Sameple1:	**21**	**24**	**25**	**26**	**27**	
Sameple2:	**22**	**27**	**28**	**30**	**31**	**36**

Can you say that the two samples came from the same population?

Solution:

Let $H_0 : \sigma_1^2 = \sigma_2^2$ and $\mu_1 = \mu_2$ (the two samples have been drawn from the same population)

$H_1 : \sigma_1^2 \neq \sigma_2^2$ and $\mu_1 \neq \mu_2$

(i) F-test : (Equality of variance)

x_1	$x_1 - \overline{x_1}$	$\left(x_1 - \overline{x_1}\right)^2$	x_2	$x_2 - \overline{x_2}$	$\left(x_2 - \overline{x_2}\right)^2$
21	-3.6	12.96	22	-7	49
24	-0.6	0.36	27	-2	4
25	0.4	0.16	28	-1	1
26	1.4	1.96	30	1	1
27	2.4	5.76	31	2	4
			36	7	49
123		21.2	174		108

$$\overline{x_1} = \frac{\sum x_1}{n_1} = \frac{123}{5} = 24.6, \overline{x_2} = \frac{\sum x_2}{n_2} = \frac{174}{6} = 29$$

$$s_1^2 = \frac{\sum\left(x - \overline{x}\right)^2}{n_1 - 1} = \frac{21.2}{4} = 5.3,\ s_2^2 = \frac{\sum\left(x_2 - \overline{x_2}\right)^2}{n_2 - 1} = \frac{108}{5} = 21.6$$

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{5(5.3)}{4} = 6.625 \quad S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{6(21.6)}{5} = 25.92$$

The test statistics is $F = \dfrac{S_2^2}{S_1^2}$ $(\sin ce\ S_1^2 < S_2^2)$

$$= \frac{25.92}{6.625} = 3.912$$

From the table, $F_{0.05}\left(n_2 - 1, n_1 - 1\right) = F_{0.05}(5,4) = 6.26$

Since $F < F_{0.05} \therefore H_0$ is accepted

(ii) <u>t-test</u>:(Equality of means)

Under H_0 , the test statistic is $t = \frac{\overline{x_1} - \overline{x_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$,

$$where\ S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{5(5.3) + 6(21.6)}{5 + 6 - 2}} = 4.164$$

$$t = \frac{24.6 - 29}{4.16\sqrt{\frac{1}{5} + \frac{1}{6}}} = -1.746 \ \Rightarrow |t| = 1.746$$

From the table ,with degrees of freedom $n_1 + n_2$ -2=9, $t_{0.05}$ =2.262

$\sin ce\ |t| < t_{0.05}\ \ \therefore H_0$ is accepted ie. $\mu_1 \neq \mu_2$

<u>Conclusion:</u> $\therefore$ The two samples could have been drawn from the same normal population.

5. **Two random samples gave the following results:**

Sample	Size	Sample mean	Sum of squares of deviations from the mean
1	**10**	**15**	**90**
2	**12**	**14**	**108**

Test whether the samples come from the same normal population at 5% level of significance.

Solution:

A normal population has 2 parameters namely mean µ and variance σ^2. To test if independent samples have been drawn from the same normal population, we have to test

1) Equality of population means using t-test or z-test, according to sample size.

2) Equality of population variances using F-test.

Given $\bar{x} = 15,\ \bar{y} = 14,\ n_1 = 10,\ n_2 = 12, \sum(x - \bar{x})^2 = 90, \sum(y - \bar{y})^2 = 108$

i) t-test to test equality of population means:

Null hypothesis H_0: $\mu_1 = \mu_2$ there is no difference between the two population means.

Alternate Hypothesis H_1: $\mu_1 \neq \mu_2$ there is difference between the two population means.

Level of Significance : $\alpha = 5\% = 0.05$ (Two tailed test)

Test statistic: $t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Where $S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum(x - \bar{x})^2 + \sum(y - \bar{y})^2\right] = \frac{1}{10 + 12 - 2}(90 + 108) = 9.9$

$S = \sqrt{9.9}$ =3.146

$$t = \frac{15-14}{3.146\sqrt{\frac{1}{10}+\frac{1}{12}}} = 0.742$$

Critical value: The critical value of t at 5% level of significance with degrees of freedom $n_1 + n_2 - 2 = 10 + 12 - 2 = 20$ is 2.086

Conclusion: calculated value < table value

H_0 is Accepted.

ii) F-test to test equality of populations variances:

Null Hypothesis H₀: $\sigma_1^2 = \sigma_2^2$ The population Variances are equal

Alternative Hypothesis H₁: $\sigma_1^2 \neq \sigma_2^2$ The population Variances are not equal

Level of significance: $\alpha = 5\%$

Test Statistics:

$$F = \frac{S_1{}^2}{S_2{}^2}$$

Where $S_1^2 = \frac{1}{n_1 - 1}\sum (x - \bar{x})^2 = \frac{1}{10-1}(90) = 10$

$$S_1^2 = \frac{1}{n_1 - 1}\sum (y - \bar{y})^2 = \frac{1}{12-1}(108) = 9.818$$

Here $S_1{}^2 > S_2{}^2 \quad \therefore F = \frac{S_1{}^2}{S_2{}^2} = \frac{10}{9.818} = 1.02$

Critical value: The critical value of *F* at 5% level of significance with degrees of freedom $(n_1 - 1, n_2 - 1) = (9,11)$ is 2.90

Here calculated value < table value, we accept H_0

Conclusion: Both null hypothesis $\mu_1 \neq \mu_2$ and $\sigma_1^2 = \sigma_2^2$ are accepted.

Hence we may conclude the two samples are drawn from same normal population.

III χ^2 **-test:**

(i). χ^2-Test for a specified population variance

(ii). χ^2-test is used to test whether differences between observed and expected frequencies are significant (goodness of fit).

(iii). χ^2-test is used to test the independence of attributes.

χ^2 **-Test for a specified population variance:**

The test statistics $\chi^2 = \frac{ns^2}{\sigma^2}$

Which follows χ^2 - distribution with (n – 1) degrees of freedom

<u>Problem:</u>

1. **The lapping process is used to grind certain silicon wafers to the proper thickness is acceptable only σ, the population S.D. of the thickness of dice cut from the wafers, is at most 0.5mil. Use the 0.05 level of significance to test the null hypothesis σ=0.5 against the alternative hypothesis σ>0.5, if the thickness of 15 dice cut from such wafers have S.D of 0.64mil.**

Solution:

Given $n = 15$, s=0.64, σ =0.5

$H_0 : \sigma = 0.5$, $H_1 : \sigma > 0.5$

Under H_0 , The test statistics $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{15\ (0.64)^2}{(0.5)^2} = 24.576$

From χ^2 table, with degrees of freedom = 14, $\chi^2_{0.05} = 23.625$

$\therefore \chi^2 > \chi^2_{0.05}$ H_0 is rejected. Hence $\sigma > 0.5$

χ^2-test is used to test whether differences between observed and expected frequencies are significant (goodness of fit):

The test statistics $\chi^2 = \sum_i \left[\frac{(O_i - E_i)^2}{O_i} \right]$

Where O_i is observed frequency, and E_i is the expected frequency.

If the data given in a series of n number, then degree of freedom = n - 1 .

Note: In case of binomial distribution d.f = n – 1, poisson distribution d.f = n – 2, normal distribution d.f = n – 3.

Problem:

1. **The following data give the number of aircraft accident that occurred during the various days of a week:**

Days :	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**
No of accidents:	**15**	**19**	**13**	**12**	**16**	**15**

Test the whether the accident are uniformly distributed over the week.

Solution:

The expected number of accident on any day $= \frac{90}{6} = 15$

Let H_0: Accidents occur uniformly over the week

H_1: Accidents not occur uniformly over the week

Days	**Observed Freqency (O_i)**	**Expected Frequency (E_i)**	$(O_i - E_i)$	$\frac{(O_i - E_i)^2}{E_i}$
Mon	15	15	0	0
Tue	19	15	4	1.066
Wed	13	15	-2	0.266
Thu	12	15	-3	0.6
Fri	16	15	1	0.066
Sat	15	15	0	0
		90		1.998

Now, $\chi^2 = \sum_i \left[\frac{(O_i - E_i)^2}{O_i} \right] = 1.998$

Here 6 observations are given, degrees of freedom = n – 1= 6 – 1 = 5

From χ^2 table, with degrees of freedom = 5, $\chi^2_{0.05} = 11.07$

$\therefore \chi^2 < \chi^2_{0.05}$ H_0 is accepted.

Conclusion: $\therefore$ Accidents occur uniformly over the week

2. **A survey of 320 families with 5 children each revealed the following distribution:**

No. of Boys:	**5**	**4**	**3**	**2**	**1**	**0**
No. of Girls:	**0**	**1**	**2**	**3**	**4**	**5**
No. of families:	**14**	**56**	**110**	**88**	**40**	**12**

Is the result consistent with the hypothesis that male and female births are equally probable?

Solution:

Let H_0: Male and female births are equally probable

H_1: Male and female births are not equally probable

Probability of male birth $= p = \frac{1}{2}$, Probability of female birth $= q = \frac{1}{2}$

The probability of x male births in a family of 5 is $p(x) = 5C_x p^x q^{5-x}, x = 0,1,2...5$

Expected number of families with x male births $= 320 \times 5C_x p^x q^{5-x}, x = 0,1,2...5$

$$= 320 \times 5C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x}$$

$$= 320 \times 5C_x \left(\frac{1}{2}\right)^5 = 10 \times 5C_x$$

The χ^2 is calculated using the following table:

No. of Boys	**Observed Freqency** **(O_i)**	**Expected Frequency** $E_i = 10 \times 5C_x$	$(O_i - E_i)$	$\frac{(O_i - E_i)^2}{E_i}$
5	14	10	4	1.6
4	56	50	6	0.72
3	110	100	10	1
2	88	100	-12	1.44
1	40	50	-10	2
0	12	10	2	0.4
Total	320	320		7.16

$$\therefore \chi^2 = 7.16$$

The tabulated value of χ^2 for n – 1 = 6 – 1 =5 degrees of freedom at 5% level of significance $= \chi^2_{0.05} = 11.07$

Since $\chi^2 < \chi^2_{0.05}$. So we accepted H_0.

Conclusion: $\therefore$ The male and female births are equally probable.

3. **Fit a poisson distribution to the following data and test the goodness of fit.**

x:	**0**	**1**	**2**	**3**	**4**	**5**	**6**
f(x):	**275**	**72**	**30**	**7**	**5**	**2**	**1**

Solution:

Mean of the given distribution $=\bar{x}=\dfrac{\sum f_i x_i}{\sum f_i}=\dfrac{189}{392}=0.482$

To fit a poisson distribution to the given data:

We take the parameter of the poisson distribution equal to the mean of the given distribution. $=\lambda=\bar{x}=0.482$

The poisson distribution is given by $P\left(X=x\right)=\dfrac{e^{-\lambda}\lambda^{x}}{x!};\ x=0,1,2...\infty$

and the expected frequencies are obtained by $f(x)=\left(\sum f_i\right)\times\dfrac{e^{-\lambda}\lambda^{x}}{x!}=392\times\dfrac{e^{-0.482}\left(0.482\right)^{x}}{x!}$

we get $f(0)=392\times\dfrac{e^{-0.482}\left(0.482\right)^{0}}{0!}=242.1,\ f(1)=392\times\dfrac{e^{-0.482}\left(0.482\right)^{1}}{1!}=116.69$

$f(3)=4.518,\ f(4)=0.544,\ f(5)=0.052\approx 0.1,\ f(6)=0.004\approx 0$

x:	0	1	2	3	4	5	6	Total
Expected Frequency:	242.1	116.69	28.12	4.518	0.544	0.052	0.004	392

H_0: The poisson distribution fit well into the data.

H_1: The poisson distribution does not fit well into the data.

The χ^2 is calculated using the following table:

x	**Observed Freqency** **(O_i)**	**Expected Frequency** $\left(E_i\right)$	$\dfrac{\left(O_i-E_i\right)^2}{E_i}$
0	275	242.1	4.471
1	72	116.7	17.122
2	30	28.1	0.128
3	**7**	**4.5**	
4	**5** } **15**	**0.5** } **5.1**	**19.218**
5	**2**	**0.1**	
6	**1**	**0**	
Total	392	392	40.939

$$\therefore\ \chi^2=40.939$$

The tabulated value of χ^2 for = 7 – 1 –1– 3 =2 degrees of freedom at 5% level of significance $=\chi^2_{0.05}=5.991$

Since $\chi^2>\chi^2_{0.05}$. So we rejected H_0.

Conclusion: $\therefore$

The Poisson distribution is not a good fit to the given data.

χ^2 -test is used to test the independence of attributes:

An attributes means a equality or characteristic. χ^2- test is used to test whether the two attributes are associated or independent. Let us consider two attributes A and B. A is divided into three classes and B is divided into three classes.

		Attribute B			
		B1	B2	B3	Total
Attribute A	A1	a_{11}	a_{12}	a_{13}	R_1
	A2	a_{21}	a_{22}	a_{23}	R_2
	A3	a_{31}	a_{32}	a_{33}	R_3
	Total	C_1	C_2	C_3	**N**

Now, under the null hypothesis H_0: The attributes A and B are independent and we calculate the expected frequency E_{ij} for varies cells using the following formula.

$$E_{ij} = \frac{R_i \times C_j}{N}\ ,\ i = 1,2,...r\ ,\ j = 1,2,...s$$

$E\left(a_{11}\right) = \frac{R_1 \times C_1}{N}$	$E\left(a_{12}\right) = \frac{R_1 \times C_2}{N}$	$E\left(a_{13}\right) = \frac{R_1 \times C_3}{N}$	R_1
$E\left(a_{21}\right) = \frac{R_2 \times C_1}{N}$	$E\left(a_{22}\right) = \frac{R_2 \times C_2}{N}$	$E\left(a_{23}\right) = \frac{R_2 \times C_3}{N}$	R_2
$E\left(a_{31}\right) = \frac{R_3 \times C_1}{N}$	$E\left(a_{32}\right) = \frac{R_3 \times C_2}{N}$	$E\left(a_{33}\right) = \frac{R_3 \times C_3}{N}$	R_3
C_1	C_2	C_3	**N**

and we compute $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$

Which follows χ^2 distribution with n = (r-1) (s-1) degrees of freedom at 5% or 1% level of significance.

1. **Calculate the expected frequencies for the following data presuming two attributes viz., conditions of home and condition of child as independent.**

	Condition of home		
		Clean	**Dirty**
Condition of Child	**Clean**	**70**	**50**
	Fair	**80**	**20**
	Dirty	**35**	**45**

Use Chi-Square test at 5% level of significance to state whether the two attributes are independent.

Solution:

Null hypothesis H_0: Conditions of home and conditions of child are independent.

Alternate hypothesis H_1: Conditions of home and conditions of child are not independent.

Level of significance: $\alpha = 0.05$

The test statistics: $\chi^2 = \sum_{i=1}^{r} \sum_{i=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Analysis:

	Condition of home			Total
		Clean	Dirty	
Condition of Child	Clean	70	50	120
	Fair	80	20	100
	Dirty	35	45	80
Total		185	115	300

$$\text{Expected Frequency } = \frac{\text{Corresponding row total} \times \text{Column total}}{\text{Grand Total}}$$

Expected Frequency for $70 = \frac{120 \times 185}{300} = 74$, Expected Frequency for $80 = \frac{100 \times 185}{300} = 61.67$,

Expected Frequency for $35 = \frac{80 \times 185}{300} = 49.33$, Expected Frequency for $50 = \frac{120 \times 115}{300} = 46$,

Expected Frequency for $20 = \frac{100 \times 115}{300} = 38.33$, Expected Frequency for $45 = \frac{80 \times 115}{300} = 30.67$

O_{ij}	E_{ij}	O_{ij} - E_{ij}	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
70	74	-4	16	$\frac{16}{74} = 0.216$
50	46	4	16	0.348
80	61.67	18.33	335.99	5.448
20	38.33	-18.33	335.99	8.766
35	49.33	-14.33	205.35	4.163
45	30.67	14.33	205.35	6.695
Total				25.636

$\therefore \chi^2 = 25.636$

$\alpha = 0.05$ Degrees of freedom = $(r-1)(c-1) = (3-1)(2-1) = 2 \quad \therefore \chi^2_\alpha = 5.991$

Conclusion:

Since $\chi^2 > \chi^2_\alpha$, we Reject our Null Hypothesis H_0. Hence, Conditions of home and conditions of child are not independent.

2. **The following contingency table presents the reactions of legislators to a tax plan according to party affiliation. Test whether party affiliation influences the reaction to the tax plan at 0.01 level of signification.**

Reaction				
Party	**In favour**	**Neutral**	**Opposed**	**Total**
Party A	**120**	**20**	**20**	**160**
Party B	**50**	**30**	**60**	**140**
Party C	**50**	**10**	**40**	**100**
Total	**220**	**60**	**120**	**400**

Solution:

Null hypothesis H_0: Party affiliation and tax plan are independent.

Alternate hypothesis H_1: Party affiliation and tax plan are not independent.

Level of significance: $\alpha = 0.05$

The test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{i=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Analysis:

Reaction				
Party	**Infavour**	**Neutral**	**Opposed**	**Total**
Party A	120	20	20	160
Party B	50	30	60	140
Party C	50	10	40	100
Total	220	60	120	**400**

E(120)= $\frac{160 \times 220}{400} = 88$; E(20)= $\frac{160 \times 60}{400} = 24$; E(20)= $\frac{160 \times 120}{400} = 48$

E(50)= $\frac{140 \times 220}{400} = 77$; E(30)= $\frac{140 \times 60}{400} = 21$; E(60)= $\frac{140 \times 120}{400} = 42$

E(50)= $\frac{100 \times 220}{400} = 55$; E(10)= $\frac{100 \times 60}{400} = 15$; E(40)= $\frac{120 \times 100}{400} = 30$

O_{ij}	E_{ij}	O_{ij} - E_{ij}	$(O_{ij}-E_{ij})^2$	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
120	88	32	1024	11.64
20	24	-4	16	0.67
20	48	-28	784	16.33
50	77	-27	729	9.47
30	21	9	81	3.86
60	42	18	324	7.71
50	55	-5	25	0.45
10	15	-5	25	1.67
40	30	10	100	3.33
Total				55.13

$\therefore \chi^2 = 55.13$

$\alpha = 0.05$ Degrees of freedom = $(r-1)(s-1) = (3-1)(3-1) = 4 \quad \therefore \chi^2_{0.05} = 13.28$

Conclusion: Since $\chi^2 > \chi^2_\alpha$, we Reject our Null Hypothesis H_0

Hence, the Party Affiliation and tax plan are dependent.

3. **From a poll of 800 television viewers, the following data have been accumulated as to, their levels of education and their preference of television stations. We are interested in determining if the selection of a TV station is independent of the level of education**

	Educational Level			
Public	**High School**	**Bachelor**	**Graduate**	**Total**
Broadcasting	**50**	**150**	**80**	**280**
Commercial Stations	**150**	**250**	**120**	**520**
Total	**200**	**400**	**200**	**800**

(i) State the null and alternative hypotheses.

(ii) Show the contingency table of the expected frequencies. (iii) Compute the test statistic.

(iv) The null hypothesis is to be tested at 95% confidence. Determine the critical value for this test.

Solution:

(i)**Null Hypothesis**: Selection of TV station is independent of level of education

Alternative Hypothesis: Selection of TV station is not independent of level of education

(ii) **Level of significance**: $\alpha = 0.05$

Educational Level				
Public	**High School**	**Bachelor**	**Graduate**	**Total**
Broadcasting	**50**	**150**	**80**	**280**
Commercial Stations	**150**	**250**	**120**	**520**
Total	**200**	**400**	**200**	**800**

To Find Expected frequency:

$$\text{Expected Frequency } = \frac{\text{Corresponding row total} \times \text{Column total}}{\text{Grand Total}}$$

$$\text{Expected Frequency for } 50 = \frac{280 \times 200}{800} = 70, \text{Expected Frequency for } 150 = \frac{280 \times 400}{800} = 140$$

$$\text{Expected Frequency for } 80 = \frac{280 \times 200}{800} = 70, \text{Expected Frequency for } 150 = \frac{520 \times 200}{800} = 130$$

$$\text{Expected Frequency for } 250 = \frac{520 \times 400}{800} = 260, \text{Expected Frequency for } 120 = \frac{520 \times 200}{800} = 130$$

The test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{i=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Analysis:

O_{ij}	E_{ij}	O_{ij} - E_{ij}	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
50	70	-20	400	5.714
150	140	10	100	0.174
80	70	10	100	1.428
150	130	20	400	3.076
250	260	-10	100	0.385
120	130	-10	100	0.769
TOTAL				11.546

(iii)Test statistic = 11.546

(iv) Critical Chi-Square = 5.991,

Conclusion: Calculated value > table value

Hence, we reject Null Hypothesis.

Large sample:

If the size of the sample n>30, then that samplw is said to be large sample. There are four important test to test the significance of large samples.

(i). Test of significance for single mean.
(ii). Test of significance for difference of two means.
(iii). Test of significance for single proportion
(iv). Test of significance for difference of two proportions.

Note:
(i). The sampling distribution of a static is approximately normal, irrespective of whether the distribution of the population is normal or not.
(ii). The sample statistics are sufficiently close to the corresponding population parameters and hence may be used to calculate the standard errors of the sampling distribution.
(iii). **Critical values for some standard LOS's (For Large Samples)**

Nature of test	**1% (0.01)** **(99%)**	**2% (0.02)** **(98%)**	**5% (0.05)** **(95%)**	**10% (0.1)** **(90%)**
Two Tailed Test	$\|z_\alpha\| = 2.58$	$\|z_\alpha\| = 2.33$	$\|z_\alpha\| = 1.96$	$\|z_\alpha\| = 1.645$
One Tailed Test (Right tailed Test)	$z_\alpha = 2.33$	$z_\alpha = 2.055$	$z_\alpha = 1.645$	$z_\alpha = 1.28$
One Tailed Test (Left tailed Test)	$z_\alpha = -2.33$	$z_\alpha = -2.055$	$z_\alpha = -1.645$	$z_\alpha = -1.28$

Problem based on Test of significance for single mean:

The test statistic $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ where $\bar{x}$ =sample mean, μ=population mean, σ = standard deviation of population, n= sample size.

Note:

If standard deviation of population is not known then the static is $z = \frac{\bar{x} - \mu}{S / \sqrt{n}}$,

where S = standard deviation of sample.

Confident Interval:
The confident interval for μ when σ is known and sampling is done from a normal population or with a large sample is $\bar{x} \mp z_\alpha \frac{\sigma}{\sqrt{n}}$

$$\Rightarrow \left(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} \right)$$

If s is known (σ is not known): $\bar{x} \mp z_\alpha \frac{s}{\sqrt{n}}$

1. **A sample of 100 students is taken from a large population, the mean height in the sample is 160cm. Can it be reasonable regarded that in the population the mean height is 165cm, and s.d. is 10cm. and find confident limit. Use an level of significance at 1%**

Solution:

Given n = 100, $\bar{x}$ =160cm, μ=165cm, σ =10cm

Let $H_0 : \mu = 165$

$H_1 : \mu \neq 165$ (two tailed test)

Under H_0 , the test statistic is $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{160 - 165}{10 / \sqrt{100}} = -5$

$\therefore |z| = -5$

From the table, $z_{0.01}$=2.58. Since $|z| > z_{0.01}$ $\therefore H_0$ is rejected. hence $\mu \neq 165$.

Confident Interval:

$$\left(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} \right) = \left(160 - 2.58 \frac{10}{\sqrt{100}}, 160 + 2.58 \frac{10}{\sqrt{100}} \right) = (157.42, 162.58)$$

2. **The mean breaking strength of the cables supplied by a manufacture is 1800 with a S.D of 100. By a new techniques in the manufacturing process, it it claimed that the breaking strength of the cable has increased. In order to test this claim, a sample of 50 cables is tested and it is found that the mean breaking strength is 1850. Can we support the claim at 1% level of significance?**

Solu:

Given n = 50, $\bar{x}$ =1850, μ=1800, σ =100

Let $H_0 : \bar{x} = \mu$

$H_1 : \bar{x} > \mu$ (one tailed test)

Under H_0 , the test statistic is $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{1850 - 1800}{100 / \sqrt{50}} = 3.535$

$\therefore |z| = 3.535$

From the table, $z_{0.01}$=2.33. Since $|z| > z_{0.01}$ $\therefore H_0$ is rejected. hence $\bar{x} > \mu$.

3. **A sample of 900 members has a mean of 3.4 cms and s.d is 2.61 cms. Is the sample from a large population of mean 3.25cm and s.d is 2.61 cms. If the population is normal and its mean is unknown find the 95% confidence limits of true mean.**

Solution:

Given $n = 900$, $\mu = 3.25$, $\bar{x} = 3.4cm$, $\sigma = 2.61$, $s = 2.61$

Null Hypothesis H₀ : Assume that there is no significant difference between sample mean and population mean. (i.e) $\mu = 3.25$

Alternative Hypothesis H₁ : Assume that there is a significant difference between sample mean and population mean. (i.e) $\mu \neq 3.25$

Level of significance : $\alpha = 5\%$

Test Statistic :

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{3.4 - 3.25}{\frac{2.61}{\sqrt{900}}} = 1.724$$

Critical value: The critical value of z for two tailed test at 5% level of significance is 1.96

Conclusion:

i.e., $z = 1.724 < 1.96 \Rightarrow$ calculated value < tabulated value

Therefore We accept the null hypothesis H$_{0.}$

i.e., The sample has been drawn from the population with mean $\mu = 3.25$

To find confidence limit:

95% confidence limits are

$$\bar{x} \mp 1.96 \frac{\sigma}{\sqrt{n}} = 3.4 \mp 1.96 \left(\frac{2.61}{\sqrt{900}} \right) = 3.4 \mp 0.1705 = (3.57, 3.2295)$$

4. **A lathe is set to cut bars of steel into lengths of 6 centimeters. The lathe is considered to be in perfect adjustment if the average length of the bars it cuts is 6 centimeters. A sample of 121 bars is selected randomly and measured. It is determined that the average length of the bars in the sample is 6.08 centimeters with a standard deviation of 0.44 centimeters.**
(i) Formulate the hypotheses to determine whether or not the lathe is in perfect adjustment.
(ii) Compute the test statistic.
(iii) What is your conclusion?

Solution:

Given $n = 121, \ \bar{x} = 6.08, \ \mu = 6, \ S = 0.44$

Null Hypothesis H$_0$: $\mu = 6$ i.e., Assume that the lathe is in perfect adjustment

Alternative Hypothesis H$_1$: $\mu \neq 6$ i.e., Assume that the lathe is not in perfect adjustment.

Level of Significance : $\alpha = 0.05$

ii) Test Statistic :

$$z = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{6.08 - 6}{\frac{0.44}{\sqrt{121}}} = \frac{0.08}{0.04} = 2$$

Table value: Table value at 5% level of significance is 1.96

iii) Conclusion:

Here calculated value > tabulated value

Hence we reject H_0.

5. **The mean life time of a sample of 100 light tubes produced by a company is found to be 1580 hours with standard deviation of 90 hours. Test the hypothesis that the mean lifetime of the tubes produced by the company is 1600 hours.**

Solution:

Given $n = 100, \ \bar{x} = 1580, \ \mu = 1600, \ S = 90$

Null Hypothesis H$_0$: $\mu = 1600$ i.e., There is no significance difference between the sample mean and population mean

Alternative Hypothesis H1: $\mu \neq 1600$ i.e., There is a significance difference between the sample mean and population mean

Level of Significance : $\alpha = 5\% = 0.05$

Test Statistic :

$$z = \frac{\overline{x}-\mu}{\frac{S}{\sqrt{n}}} = \frac{1580-1600}{\frac{90}{\sqrt{100}}} = \frac{-20}{9} = -2.22$$

$$|z| = 2.22$$

Table value: Table value at 5% level of significance is 1.96 (two tailed test)

Conclusion:

Here calculated value > tabulated value

Hence we reject H_0.

Hence the mean life time of the tubes produced by the company may not be 1600 hrs.

Problem based on Test of significance for difference of two means:

The test statistic $z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ where σ_1, σ_2 are S.D. of populations.

Test Statistic:

i) $Z = \frac{\overline{x_1} - \overline{x_2}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ If σ is known and $\sigma_1 = \sigma_2$

ii) $Z = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ If σ is not known and $\sigma_1 \neq \sigma_2$, S_1^2, S_2^2 are known.

Confident Interval:

The confident interval for difference between two population mean for large sample,

(1) when $\sigma(\sigma_1, \sigma_2)$ is known is $(\overline{x}_1 - \overline{x}_2) \pm z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

(2). when s (s_1, s_2) is known is $(\overline{x}_1 - \overline{x}_2) \pm z_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

1. **In a random sample of size 500, the mean is found to be 20. In another independent sample of size 400, the mean is 15. Could the samples have been drawn from the same population with S.D 4?**

Solution:

Given $\overline{x}_1 = 20,\ \overline{x}_2 = 15,\ n_1 = 500,\ n_2 = 400, \sigma = 4$

Null hypothesis H_0: $\mu_1 = \mu_2$ The samples have been drawn from the same population.

Alternate Hypothesis H_1: $\mu_1 < \mu_2$ The samples could not have been drawn from same population.

Level of Significance : $\alpha = 5\% = 0.05$ (Two tailed test)

Test statistic: $z = \dfrac{\overline{x_1} - \overline{x_2}}{\sigma\sqrt{\dfrac{1}{n_2} + \dfrac{1}{n_1}}} = \dfrac{20-15}{4\sqrt{\dfrac{1}{500} + \dfrac{1}{400}}} = 18.6$

Critical value: The critical value of t at 1% level of significance is 2.58
Conclusion: calculated value > table value

H_0 is rejected

The samples could not have been drawn from same population.

2. **Test significance of the difference between the means of the samples, drawn from two normal populations with the same SD using the following data:**

	Size	Mean	Standard Deviation
Sample I	**100**	**61**	**4**
Sample II	**200**	**63**	**6**

Solution:

Given $\bar{x}_1 = 60,\ \bar{x}_2 = 63, s_1 = 4,\ s_2 = 6, n_1 = 100,\ n_2 = 200$

Null hypothesis $H_0:\ \mu_1 = \mu_2$ there is no significance difference between the means of the samples.

Alternate Hypothesis $H_1:\ \mu_1 \neq \mu_2$ there is a significance difference between the means of the samples.

Level of Significance : $\alpha = 5\% = 0.05$ (two tailed test)

Test statistic: $z = \dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\dfrac{{s_1}^2}{n_2} + \dfrac{{s_2}^2}{n_1}}} = \dfrac{61-63}{\sqrt{\dfrac{4^2}{200} + \dfrac{6^2}{100}}} = -3.02 \Rightarrow |z| = 3.02$

Critical value: The critical value of t at 5% level of significance is 1.96
Conclusion: calculated value > table value

H_0 is rejected .Therefore the two normal populations, from which the samples are drawn, may not have the same mean though they may have the same S.D.

3. **A sample of heights of 6400 Englishmen has a mean of 170cm and a S.D of 6.4cm, while a simple sample of heights of 1600 Americans has a mean of 172cm and a S.D of 6.3cm. D the data indicate that Americans are on the average, taller than Englishmen?**

Solution:

Given $\bar{x}_1 = 170,\ \bar{x}_2 = 172, s_1 = 6.4,\ s_2 = 6.3,\ n_1 = 6400,\ n_2 = 1600$

Null hypothesis $H_0:\ \mu_1 = \mu_2$ there is no significance difference between the heights of Americans and Englishmen.

Alternate Hypothesis $H_1:\ \mu_1 < \mu_2$ Americans are on the average, taller than Englishmen

Level of Significance : $\alpha = 5\% = 0.05$ (one tailed test)

Test statistic: $z = \dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\dfrac{{s_1}^2}{n_2} + \dfrac{{s_2}^2}{n_1}}} = \dfrac{170-172}{\sqrt{\dfrac{6.4^2}{6400} + \dfrac{6.3^2}{1600}}} = -11.32 \Rightarrow |z| = 11.32$

Critical value: The critical value of t at 5% level of significance is 1.645

Conclusion: calculated value > table value

H_0 is rejected. We conclude that the data indicate that Americans are on the average, taller than Englishmen.

4. **The aveage marks scored by 32 boys is 72 with a S.D of 8, while that for 36 girls is 70 with a S.D of 6. Test at 1%level of significance whether the boys perform beter than girls.**

Solution:

Given $\bar{x}_1 = 72,\ \bar{x}_2 = 70, s_1 = 8,\ s_2 = 6,\ n_1 = 32,\ n_2 = 36$

$H_0:\ \mu_1 = \mu_2$ (Both perfom are equal)

$H_0:\ \mu_1 > \mu_2$ (Boys are better than girls) (one tailed test)

The test statistic: $z = \dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\dfrac{s_1{}^2}{n_2} + \dfrac{s_2{}^2}{n_1}}} = \dfrac{72-70}{\sqrt{\dfrac{8^2}{32} + \dfrac{6^2}{36}}} = 1.15$

Critical value: The critical value of t at 1% level of significance is 2.33

Conclusion: calculated value < table value

H_0 is accepted. Hence both are equal.

Problem based on Test of significance for single proportion:

To test the significant difference between the sample proportion p and the population proportion P, then we use the test statistic

$z = \dfrac{p-P}{\sqrt{\dfrac{PQ}{n}}}$, where Q = 1 – P

Confident Interval:

The confident interval for population proportion for large sample is $p \mp z_\alpha \sqrt{\dfrac{PQ}{n}}$

1. **In a big city 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers?**

Solution:

Given n=600 , Number of smokers=325

p = sample proportion of smokers $\Rightarrow p = 325/600 = 0.5417$

P= Population proportion of smokers in the city $= 1/2 = 0.5 \Rightarrow Q = 0.5$

Null Hypothesis H0: The number of smokers and non-smokers are equal in the city.

Alternative Hypothesis H1: P > 0.5 (Right Tailed)

Test Statistic:

$$z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.5417-0.5}{\sqrt{\frac{0.5*0.5}{600}}} = 2.04$$

Critical value:
Tabulated value of z at 5% level of significance for right tail test is 1.645.

Conclusion:

Since Calculated value of z > tabulated value of z.

We reject the null hypothesis. The majority of men in the city are smokers.

2. **40 people were attacked by a disease and only 36 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease, is 85% at 5% level of significance?**

Solution:
Given

The Sample proportion, $p = \frac{36}{40} = 0.90$

Population proportion $P = 0.85 \Rightarrow Q = 1 - P = 1 - 0.85 = 0.15$

Null Hypothesis H₀: $P = 0.85$ i.e., There is no significance difference in survival rate

Alternative Hypothesis H₁: $P \neq 0.85$

i.e., There is a significance difference in survival rate.

Level of Significance : $\alpha = 0.05$

Test Statistic :

$$z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.90-0.85}{\sqrt{\frac{0.85\times 0.15}{40}}} = 0.886$$

Table value: Tabulated value of z at 5% level of significance is 1.96

Conclusion : The table value >calculated value
Hence we accept the null hypothesis
Conclude that the survival rate may be taken as 85%.

3. **A Manufacturer of light bulbs claims that an average 2% of the bulbs manufactured by his firm are defective. A random sample of 400 bulbs contained 13 defective bulbs. On the basis of this sample, can you support the manufacturer's claim at 5% level of significance?**

Solution:
Given $n = 400$

$p =$ Sample proportion of defectives = $\frac{X}{n} = \frac{13}{400} = 0.0325$

Null Hypothesis H₀: $P = 2\% = 0.02$ i.e., Assume that 2% bulbs are defective.

Alternative Hypothesis H₁: $P \neq 2\% \neq 0.02$ i.e., Assume that 2% bulbs are non-defective.

Level of significance: $\alpha = 5\% = 0.05$

Test Statistic : $z = \frac{p-P}{\sqrt{\frac{PQ}{n}}}$

$$z = \frac{0.0325-0.02}{\sqrt{\frac{0.02\times 0.98}{400}}} = \frac{0.0125}{0.0007} = 1.7857$$

Critical value : The critical value of tat 5% level of significance is 1.645 (one tailed test)

Conclusion:

Here calculated value > table value.

So we accept H_0 . Hence the manufacturers claim cannot be supported.

4. **A salesman in a departmental store claims that at most 60 percent of the shoppers entering the store leave without making a purchase. A random sample of 50 shoppers should that 35 out of them left without making a purchase. Are these sample reults consistent with the claim of the salesman? Use an LOS of 0.05.**

Solution:

Let p = Sample proportion of shoppers not making a purchase = $\frac{35}{50} = 0.7$

P = Population proportion of shoppers not making a purchase = $60\% = \frac{60}{100} = 0.6$,

and Q = 1 – P = 0.4

H0: $P = 0.6$ i.e., The claim is accepted

H1: $P \neq 0.6$ (two tailed test)

The test Statistic is $z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.7-0.6}{\sqrt{\frac{0.6\times0.4}{50}}} = 1.445$

From the table, $z_{0.05}$=1.96. Since $|z| < z_{0.05} \therefore H_0$ is accepted

Conclusion:

The sample reults are consistent with the claim of the salesman.

Problem based on Test of significance for Two proportion:

To test the significant difference between the sample proportion p_1 and p_2 and the population proportion P, then we use the test statistic

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$, where Q = 1 – P

If P is not known, then $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

Confident Interval:

The confident interval for difference between two population proportion for large sample is $\left(p_1 - p_2\right) \mp z_\alpha \sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

1. **Before an increase in excise duty on tea, 800 people out of a sample of 1000 were consumers of tea. After the increase in duty, 800 people were consumers of tea in a sample of 1200 persons. Find whether there is significant decrease in the consumption of tea after the increase in duty. Also find confident limit.**

Solution:

Given $n_1 = 1000, n_2 = 1200$

$p_1 = \text{proportion of tea drinkers before increase inexcise duty} = \frac{800}{1000} = 0.8$

$p_2 = \text{proportion of tea drinkers before increase inexcise duty} = \frac{800}{1200} = 0.6667$

Null hypothesis: $H_0 : P_1 = P_2$ there is no significance difference in the consumption of tea before after increase in excise duty

Alternate hypothesis: $H_1 : P_1 \neq P_2$ there is a significance difference in the consumption of tea before after increase in excise duty

Level of significance: $\alpha = 5\%$ =0.05

Test Statistic: $z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

Where

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(0.8)(1000) + (0.67)(1200)}{1000 + 1200} = 0.7273 \quad \Rightarrow Q = 1 - P = 1 - 0.7273 = 0.2727$$

$$z = \frac{0.8 - 0.6667}{\sqrt{(0.7273)(0.2727)\left(\frac{1}{1000} + \frac{1}{1200}\right)}} = \frac{0.1333}{0.01907} = 6.99$$

Critical value: the critical value of z at 5% level of significance is 1.645

Conclusion:

Here calculated value > table value

$\therefore$ We reject H_0

Hence there is no significance difference in the consumption of tea before after increase in excise duty.

Confident Interval:

The confident interval for difference between two population proportion for large sample is

$$\left(p_1 - p_2\right) \mp z_\alpha \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \left[\left(0.8 - 0.667\right) \mp 1.645 \sqrt{0.7273 \times 0.2727\left(\frac{1}{1000} + \frac{1}{1200}\right)}\right]$$

$$= (0.1016, 0.1644)$$

2. **Random samples of 400 men and 600 women asked whether they would like to have a flyover near their residence.200 men and 325 women were in favor of the proposal. Test the hypothesis that proportions of men and women in favor of the proposal are same against that they are not, at 5% level.**

Solution:

Given $n_1 = 400, n_2 = 600$

$p_1 = \text{proportion of men} = \frac{200}{400} = 0.5$

$p_2 = \text{proportion of women} = \frac{325}{600} = 0.541$

Null hypothesis: $H_0 : P_1 = P_2$ Assume that there is no significant difference between the option of men and women as far as proposal of flyover is concerned.

Alternate hypothesis: $H_1 : P_1 \neq P_2$ Assume that there is significant difference between the option of men and women as far as proposal of flyover is concerned

Level of significance: $\alpha = 5\%$ =0.05 (two tailed)

Test Statistic: $z = \dfrac{p_1 - p_2}{\sqrt{PQ\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

Where $P = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \dfrac{(400)(0.5) + (600)(0.541)}{400 + 600} = 0.525 \quad \Rightarrow Q = 1 - P = 1 - 0.525 = 0.475$

$$z = \frac{0.5 - 0.541}{\sqrt{(0.525)(0.475)\left(\dfrac{1}{400} + \dfrac{1}{600}\right)}} = \frac{-0.041}{0.032} = -1.34 \quad \Rightarrow |z| = 1.34$$

Critical value: the critical value of z at 5% level of significance is 1.96

Conclusion:

Here calculated value < table value

$\therefore$ We accept H_0 at 5% level of significance.

Hence There is no difference between the option of men and women as far as proposal of flyover are concerned.

3. **A machine puts out 16 imperfect articles in a sample of 500. After the machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved?**

Solution:

Hypothesis:

$H_0 : P_1 = P_2$

$H_1 : P_1 > P_2$

Level of Significance : $\alpha = 0.05$

Test Statistic : $Z = \dfrac{p_1 - p_2}{\sqrt{PQ\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

Analysis:

The Sample proportion,

$p_1 = \dfrac{16}{500} = 0.032, \quad p_2 = \dfrac{3}{100} = 0.03, \quad P = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0.032$ & $Q = 1 - P = 0.968$

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \frac{0.032 - 0.03}{\sqrt{0.032 \times 0.968\left(\dfrac{1}{500} + \dfrac{1}{100}\right)}} = 0.1037$$

Table value : $Z_\alpha = 1.645$

Conclusion:

Calculated value < table value

Hence we accept the null hypothesis and conclude that the machine has not improved after overhauling.

UNIT-IV DESIGN OF EXPERIMENT

The sequence of steps taken to ensure a scientific analysis leading to valid inferences about the hypothesis is called design of experiment. The main aim of the design of experiments is to control the extraneous variables and hence to minimize the experimental error so that the results of the experiments could be attributed only to the experimental variables.

The basic principles of design of experiments:

(i) Randomization

(ii) Replication

(iii) Local Control

Basic design of Experiments:

Depending on the number of extraneous variables whose effects are to be controlled, various design procedures are developed in the study of experimental design. We shall consider here three important designs.

(1) Completely randomized Design (C.R.D)

(2) Randomized Block Design (R.B.D)

(3) Latin Square Design (LSD)

ANOVA:

Analysis of Variance is a technique that will enable us to test for the significance of the difference among more than two sample means.

Assumptions of analysis of variance:

(i) The sample observations are independent
(ii) The environmental effects are additive in nature
(iii) The samples have been randomly selected from the population.
(iv) Parent population from which observations are taken in normal.

One Way Classification (or) Completely randomized Design (C.R.D)

The C.R.D is the simplest of all the designs, based on principles of randomization and replication. In this design, treatments are allocated at random to the experimental units over the entire experimental materials.

Advantages of completely randomized block design:

The advantages of completely randomized experimental design as follows:

(i) Easy to lay out.
(ii) Allow flexibility
(iii) Simple statistical analysis
(iv) lots of information due to missing data is smaller than with any other design

Working Procedure (One – Way classification)

Null Hypothesis H_0: There is no significance difference between the treatments.

Alternate Hypothesis H_1: There is a significance difference between the treatments.

Analysis:

Step 1: Find *N*= number of observations

Setp 2: Find *T* = The total value of observations

Step 3: Find the correction Factor = $C.F = \frac{T^2}{N}$

Step 4: Calculate the total sum of squares = $TSS = \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \ldots\right) - C.F$

Step 4: Find Total Sum of Square $TSS = \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \ldots\right) - C.F$

Step 5: Column Sum of Square $SSC = \left(\frac{\left(\sum X_1\right)^2}{N_1} + \frac{\left(\sum X_2\right)^2}{N_2} + \frac{\left(\sum X_3\right)^2}{N_3} + \ldots\right) - C.F$

Where N_i = Total number of observation in each column ($i = 1, 2, 3, \ldots$)

Step 6: Prepare the ANOVA TABLE to calculate F-ratio.

Source of Variation	Sum of Degrees	Degree of freedom	Mean Square	F- Ratio
Between Columns	SSC	c-1	$MSC = \frac{SSC}{c-1}$	$F_C = \frac{MSC}{MSE}$ if $MSC > MSE$ (or) $F_C = \frac{MSE}{MSC}$ if $MSE > MSC$
Error	SSE	N-c	$MSE = \frac{SSE}{N-c}$	
Total	SST	N-1		

Step 7: Find the table value (use χ^2 table)

Step 8: Conclusion:

Calculated value < Table Value, the we accept Null Hypothesis H_0 **(or)**

Calculated value > Table Value, the we reject Null Hypothesis H_0

1. **The following are the number of mistakes made in 5 successive days by 4 technicians working for a photographic laboratory. Test whether the difference among the four sample means can be attributed to chance. (Test at a level of significance $\alpha = 0.01$)**

Technicians			
I	**II**	**III**	**IV**
6	**14**	**10**	**9**
14	**9**	**12**	**12**
10	**12**	**7**	**8**
8	**10**	**15**	**10**
11	**14**	**11**	**11**

Solution:

H_0: There is no significant difference between the technicians

H_1 : Significant difference between the technicians

We shift the origin to 10

	X_1	**X_2**	**X_3**	**X_4**	**TOTAL**	**X_1^2**	**X_2^2**	**X_3^2**	**X_4^2**
Total	-4	4	0	-1	**-1**	16	16	0	1
	4	-1	2	2	**7**	16	1	4	4
	0	2	-3	-2	**-3**	0	4	9	4
	-2	0	5	0	**3**	4	0	25	0
	1	4	1	1	**7**	1	16	1	1
	-1	**9**	**5**	**0**	**13**	**37**	**37**	**39**	**10**

Step1: N= Total No of Observations = 20

Step 2: T=Grand Total = 13

Step 3: Correction Factor = $\frac{(\text{Grand total})^2}{\text{Total No of Observations}} = \frac{T^2}{N} = \frac{13^2}{20} = 8.45$

Step 4: $\text{TSS} = \sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2 - C.\text{F} = 37 + 37 + 39 + 10 - 8.45 = 114.55$

Step 5:

$$\text{SSC} = \frac{\left(\sum X_1\right)^2}{N_1} + \frac{\left(\sum X_2\right)^2}{N_1} + \frac{\left(\sum X_3\right)^2}{N_1} + \frac{\left(\sum X_4\right)^2}{N_1} - C.\text{F}$$

$$= \frac{(-1)^2}{5} + \frac{9^2}{5} + \frac{5^2}{5} + 0 - 8.45$$

$SSC = 0.2 + 16.2 + 5 - 8.45 = 12.95$

Where N_1 = Number of elements in each column=5

Step 7: $\text{SSE=TSS-SSC} = 114.5 - 12.95 = 101.6$

Step 8: **ANOVA TABLE:**

Source of Variation	**Sum of Squares**	**Degree of freedom**	**Mean Square**	**F- Ratio**
Between Columns	SSC=12.95	C-1= 4-1=3	$MSC = \frac{SSC}{C-1}$ =4.317	$F_C = \frac{MSC}{MSE}$ $= \frac{6.35}{4.317}$ =1.471
Error	SSE=101.6	N-C=20-4=16	$MSE = \frac{SSE}{N-C}$ =6.35	

Cal F_C = 1.471

Table value : F_C (16,3)=5.29

Conclusion : Cal F_C< Tab F_C

⇒ There is no significance difference between the technicians

2. **A completely randomized design exprement with 10 plots and 3 treatments gave the following results.**

Plot No :	1	2	3	4	5	6	7	8	9	10
Treatment :	A	B	C	A	C	C	A	B	A	B
Yield :	5	4	3	7	5	1	3	4	1	7

Analyse the results for treatment effects.

Solution:

A	***B***	***C***
5	***4***	***3***
7	***4***	***5***
3	***7***	***1***
1		

Null Hypothesis H_0: *There is no significant difference in treatments*

Alternate Hypothesis H_1 : *Significant difference in treatments*

	X_1	***X_2***	***X_3***	***TOTAL***	***X_1^2***	***X_2^2***	***X_3^2***
Total	*5*	*4*	*3*	***12***	*25*	*16*	*9*
	7	*4*	*5*	***16***	*49*	*16*	*25*
	3	*7*	*1*	***11***	*9*	*49*	*1*
	1			***1***	*1*		
	16	*15*	*9*	***40***	*84*	*81*	*35*

Step1: N= Total No of Observations = 10

Step 2: T=Grand Total = 40

Step 3: Correction Factor = $\frac{(\text{Grand total})^2}{\text{Total No of Observatio ns}} = \frac{T^2}{N} = \frac{40^2}{10} = 160$

Step 4: $\text{TSS} = \sum X_1^2 + \sum X_2^2 + \sum X_3^2 - C.\text{F} = 84 + 81 + 35 - 160 = 40$

Step 5: $\text{SSC} = \frac{\left(\sum X_1\right)^2}{N_1} + \frac{\left(\sum X_2\right)^2}{N_1} + \frac{\left(\sum X_3\right)^2}{N_1} - C.\text{F} = \frac{(16)^2}{4} + \frac{15^2}{3} + 3 - 160$

$SSC = 64 + 75 + 27 - 160 = 6$

Where N_1 *= Number of elements in each column*

Step 7: $\text{SSE=TSS-SSC} = 40 - 6 = 34$

Step 8: ***ANOVA TABLE:***

Source of Variation	***Sum of Squares***	***Degree of freedom***	***Mean Square***	***F- Ratio***
Between Columns	*SSC=6*	*C-1= 3-1=2*	$\text{MSC} = \frac{\text{SSC}}{\text{C} - 1}$ $= \frac{6}{2} = 3$	$\text{F}_\text{C} = \frac{\text{MSE}}{\text{MSC}}$ $= \frac{4.86}{3}$ $= 1.62$
Error	*SSE=34*	*N-C=10-3=7*	$\text{MSE} = \frac{\text{SSE}}{\text{N} - \text{C}}$ $= \frac{34}{87} = 4.86$	

Cal F_C *= 1.62*

Table value : *F_C (7,2)=19.35*

Conclusion : *Cal F_C< Tab F_C*

We accept Null Hypothesis ⇒ There is no significance difference in tretments

3. **As head of the department of a consumers research organization you have the responsibility of testing
and comparing life times of 4 brands of electric bulbs.suppose you test the life time of 3 electric bulbs
each of 4 brands,the data is given below,each entry representing the life time of an electric bulb,measured
in hundreds of hours**

A	B	C	D
20	**25**	**24**	**23**
19	**23**	**20**	**20**
21	**21**	**22**	**20**

<u>Solution:</u>

H_0: Here the population means are equal.

H_1: The population mean are not equal.

	X_1	X_2	X_3	X_4	X_1^2	X_2^2	X_3^2	X_4^2
	20	**25**	**24**	**23**	400	625	576	529
	19	**23**	**20**	**20**	361	529	400	400
	21	**21**	**22**	**20**	441	441	484	400
TOTAL	60	69	66	63	1202	1595	1460	1329

Step1: N= Total No of Observations = 12

Step 2: T=Grand Total = 258

Step 3: Correction Factor = $\frac{(\text{Grand total})^2}{\text{Total No of Observations}} = \frac{T^2}{N} = \frac{258^2}{12} = 5547$

Step 4: $\text{TSS} = \sum X_1^2 + \sum X_2^2 + \sum X_3^2 - C.\text{F} = 39$

Step 5: $\text{SSC} = \frac{\left(\sum X_1\right)^2}{N_1} + \frac{\left(\sum X_2\right)^2}{N_1} + \frac{\left(\sum X_3\right)^2}{N_1} - C.\text{F} = 15$

Where N_1 *= Number of elements in each column*

Step 7: $\text{SSE=TSS-SSC} = 39 - 15 = 24$

Step 8: ***ANOVA TABLE:***

Source of Variation	***Sum of Squares***	***Degree of freedom***	***Mean Square***	***F- Ratio***
Between Columns	*SSC=39*	*C-1= 4-1=3*	$\text{MSC} = \frac{\text{SSC}}{\text{C}-1}$ *=13*	$\text{F}_\text{C} = \frac{\text{MSE}}{\text{MSC}}$ $= \frac{13}{1.875}$ $= 6.93$
Error	*SSE=15*	*N-C=12-4=8*	$\text{MSE} = \frac{\text{SSE}}{\text{N}-\text{C}}$ *=1.875*	

Cal F$_\text{C}$ = 6.93 & Tab F$_\text{C}$ (3,8)=4.07

Conclusion : Cal F$_\text{C}$ **>** Tab F$_\text{C}$ Hence we rejected **H0**

Two Way Classification (or) Randomized Block Design (R.B.D):

The entire experiment influences on only two factors is two way Classification.

Working Procedure (Two – Way classification)

Null Hypothesis H_0: There is no significance difference between the treatments.

Alternate Hypothesis H_1: There is a significance difference between the treatments.

Analysis:

Step 1: Find N= number of observations

Setp 2: Find T = The total value of observations

Step 3: Find the correction Factor = $C.F = \frac{T^2}{N}$

Step 4: Calculate the total sum of squares = $\text{TSS} = \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \ldots\right) - C.\text{F}$

Step 4: Find Total Sum of Square $\text{TSS} = \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \ldots\right) - C.\text{F}$

Step 5: Find column sum of Square $\text{SSC} = \left(\frac{\left(\sum X_1\right)^2}{N_1} + \frac{\left(\sum X_2\right)^2}{N_2} + \frac{\left(\sum X_3\right)^2}{N_3} + \ldots\right) - C.\text{F}$

Where N_i = Total number of observation in each column $(i = 1, 2, 3, \ldots)$

Step 6: Find Row sum of square = $SSR = \left(\frac{(\sum Y_1)^2}{N_1} + \frac{(\sum Y_2)^2}{N_2} + \frac{(\sum Y_3)^2}{N_3} + \ldots \right) - C.F$

Where N_j = Total number of observation in each Row $(j = 1, 2, 3, \ldots)$

Step 7: Prepare the ANOVA TABLE to calculate F-ratio.

Source of Variation	Sum of Degrees	Degree of freedom	Mean Square	F- Ratio
Between Columns	SSC	c-1	$MSC = \frac{SSC}{c-1}$	$F_C = \frac{MSC}{MSE}$ if $MSC > MSE$ (or) $F_C = \frac{MSE}{MSC}$ if $MSE > MSC$
Between Rows	SSR	r-1	$MSC = \frac{SSR}{r-1}$	$F_R = \frac{MSR}{MSE}$ if $MSR > MSE$ (or) $F_R = \frac{MSE}{MSR}$ if $MSE > MSR$
Error	SSE	N-c-r+1	$MSE = \frac{SSE}{N-c-r+1}$	
Total	TSS	rc-1		

Step 8: Find the table value for both F_C & F_R (use χ^2 table)

Step 9: Conclusion:

Calculated value < Table Value, the we accept Null Hypothesis H_0 **(or)**

Calculated value > Table Value, the we reject Null Hypothesis H_0

1. **A Company appointments four salesmen A, B, C and D and observes their sales in 3 seasons: summer, winter and monsoon. The figures (in lakhs of Rs.) are given in the following table:**

Season	**Salesman**			
	A	**B**	**C**	**D**
Summer	**36**	**36**	**21**	**35**
Winter	**28**	**29**	**31**	**32**
Monsoon	**26**	**28**	**29**	**29**

i) Do the salesmen significantly differ in performance?

ii) Is there significant difference between the seasons?
Solution:

Null Hypothesis H_0**:** There is no significant difference between the sales in the 3 seasons and also between the sales of the 4 salesmen.

Alternate Hypothesis H_1**:** There is a significant difference between the sales in the 3 seasons and also between the sales of the 4 salesmen.

Test statistic:
To simplify calculations we deduct 30 from each value

Seasons		**A** **X_1**	**B** **X_2**	**C** **X_3**	**D** **X_4**	**Seasons** **Total**	**X_1^2**	**X_2^2**	**X_3^2**	**X_4^2**
Y_1	Summer	6	6	-9	5	**8**	36	36	81	25
Y_2	Winter	-2	-1	1	2	**0**	4	1	1	4
Y_3	Monson	-4	-2	-1	-1	**-8**	16	4	1	1
Total		0	3	-9	6	**0**	56	41	83	30

Step1: N= Total No of Observations = 12

Step 2: T=Grand Total = 0

Step 3: Correction Factor = $\frac{(\text{Grand total})^2}{\text{Total No of Observations}} = \frac{T^2}{N} = \frac{0^2}{12} = 0$

Step 4: $\text{TSS} = \sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2 - C.\text{F} = 56 + 41 + 83 + 30 - 0 = 210$

Step 5: $\text{SSC} = \frac{\left(\sum X_1\right)^2}{N_1} + \frac{\left(\sum X_2\right)^2}{N_1} + \frac{\left(\sum X_3\right)^2}{N_1} + \frac{\left(\sum X_4\right)^2}{N_1} - C.\text{F} = \frac{0^2}{3} + \frac{3^2}{3} + \frac{(-9)^2}{3} + \frac{6^2}{3} - 0$

$\text{SSC} = 0 + 3 + 27 + 12 - 0 = 42$

Where N_1 = Number of elements in each column

Step 6: $\text{SSR} = \frac{\left(\sum Y_1\right)^2}{N_2} + \frac{\left(\sum Y_2\right)^2}{N_2} + \frac{\left(\sum Y_3\right)^2}{N_2} - C.\text{F} = \frac{8^2}{4} + \frac{0^2}{4} + \frac{(-8)^2}{4} + \frac{6^2}{4} - 0 = 16 + 0 + 16 - 0 = 32$

Where N_2 = Number of elements in each row

Step 7: $\text{SSE=TSS-SSC-SSR} = 210 - 42 - 32$

Step 8: **ANOVA TABLE:**

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Sum of Squares**	**varience**	**F – ratio**
Between Columns (Salesmen)	SSC=42	c-1=4-1=3	$MSC = \frac{SSC}{c-1}$ $= \frac{42}{3} = 14$	$MSC = \frac{MSE}{MSC}$ $= \frac{22.67}{14}$ $= 1.619$	$F_C(6,3) = 8.94$
Between rows (Seasons)	SSR =32	r-1=3-1=2	$MSR = \frac{SSR}{r-1}$ $= \frac{32}{2} = 16$	$MSR = \frac{MSE}{MSR}$ $= \frac{22.67}{16}$ $= 1.417$	$F_R(6,2) = 8.94$
Error	SSE=136	N-c-r +1=6	$MSE = \frac{SSE}{N-c-r+1}$ $= \frac{136}{6} = 22.67$		
Total	210	11			

Table Value of F = $F_C(\text{Error, d.f}) = F_C(6,3) = 8.94$, $F_R(\text{Error, d.f}) = 8.94$ with 5% level of significance

Conclusion:

1) Cal F_R < Table $F_{R,0.05}(6,3)$

Hence we accept the H_0 and we conclude that there is no significant difference between sales in the three seasons.

2) **Cal F_R < Table $F_{R,0.05}(6,2)$** .

Hence we accept the H_0 and we conclude that there is no significant difference between in the sales of 4 salesmen.

2. **The following data represent the number of units of production per day turned out by different workers using 4 different types of machines.**

		Machine type			
		A	**B**	**C**	**D**
Workers	**1**	**44**	**38**	**47**	**36**
	2	**46**	**40**	**52**	**43**
	3	**34**	**36**	**44**	**32**
	4	**43**	**38**	**46**	**33**
	5	**38**	**42**	**49**	**39**

(1) Test whether the five men differ with respect to mean productivity and

(2) Test whether the mean productivity is the same for the four different machine types.

Solution:

Null Hypothesis H$_0$: There is no significant difference between the Machine types the Workers.

Alternate Hypothesis H₁ : Significant difference between the Machine types between the Workers

Test statistic:

To simplify calculations we deduct 46 from each value

	Machine type				**workers Total**	**X_1^2**	**X_2^2**	**X_3^2**	**X_4^2**
workers	**A** **X_1**	**B** **X_2**	**C** **X_3**	**D** **X_4**					
Y_1	-2	-8	1	-10	**-19**	4	64	1	100
Y_2	0	-6	6	-3	**-3**	0	36	36	9
Y_3	-12	-10	-2	-14	**-38**	144	100	4	196
Y_4	-3	-8	0	-13	**-24**	9	64	0	169
Y_5	-8	-4	3	-7	**-16**	64	16	9	49
Total	-25	-36	8	-47	**-100**	**221**	**280**	**50**	**523**

Step1: N= Total No of Observations = 20

Step 2: T=Grand Total = -100

Step 3: Correction Factor = $\frac{(\text{Grand total})^2}{\text{Total No of Observations}} = \frac{T^2}{N} = \frac{(-100)^2}{20} = \frac{10000}{20} = 500$

Step 4: $TSS = \sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2 - C.F = 221 + 280 + 50 + 523 - 500 = 574$

Step 5:

$$SSC = \frac{\left(\sum X_1\right)^2}{N_1} + \frac{\left(\sum X_2\right)^2}{N_1} + \frac{\left(\sum X_3\right)^2}{N_1} + \frac{\left(\sum X_4\right)^2}{N_1} - C.F = \frac{(-25)^2}{5} + \frac{(-36)^2}{5} + \frac{8^2}{5} + \frac{(-47)^2}{5} - 500$$

$$SSC = \frac{625}{5} + \frac{1296}{5} + \frac{64}{5} + \frac{2209}{5} - 500 = 838.8 - 500 = 338.8$$

Where N_1 = Number of elements in each column

Step 6: $SSR = \frac{\left(\sum Y_1\right)^2}{N_2} + \frac{\left(\sum Y_2\right)^2}{N_2} + \frac{\left(\sum Y_3\right)^2}{N_2} + \frac{\left(\sum Y_4\right)^2}{N_2} + \frac{\left(\sum Y_5\right)^2}{N_2} - C.F$

$$= \frac{(-19)^2}{4} + \frac{(-3)^2}{4} + \frac{(-38)^2}{4} + \frac{(-24)^2}{4} + \frac{(-16)^2}{4} - 500$$

$$= \frac{361 + 9 + 1444 + 576 + 256}{4} - 500 = 661.5 - 500 = 161.5$$

$SSR = 161.5$

Where N_2 = Number of elements in each row=4

Step 7: $SSE = TSS - SSC - SSR = 574 - 338.8 - 161.5 = 73.7$

Step 8: **ANOVA TABLE:**

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Sum of Squares**	**varience**	**F – ratio**
Between Columns (Salesmen)	SSC=338.8	c-1=4-1=3	$MSC = \frac{SSC}{c-1}$ $= \frac{338.8}{3}$ $= 112.9$	$MSC = \frac{MSC}{MSE}$ $= \frac{112.9}{6.14}$ $= 18.39$	$F_C(3,12) = 3.49$
Between rows (Seasons)	SSR =161.8	r-1=5-1=4	$MSR = \frac{SSR}{r-1}$ $= \frac{161.5}{4}$ $= 40.4$	$MSR = \frac{MSR}{MSE}$ $= \frac{40.4}{6.14}$ $= 6.58$	$F_R(4,12) = 3.26$
Error	SSE=73.7	N-c-r +1 =20-4-5+1 =12	$MSE = \frac{SSE}{N-c-r+1}$ $= \frac{73.7}{12}$ $= 6.14$		
Total	574.3	19			

Table Value of F = $F_C(\text{Error}, \text{d.f}) = F_C(6,3) = 8.94$,$F_R(\text{Error}, \text{d.f}) = 8.94$ with 5% level of significance

Conclusion:

1) Calculated value > table value

Hence we reject the H_0 and we conclude that there is significant difference between sales in the three seasons.

2) **Calculated value > table value**

Hence we rejectthe H_0 and we conclude that there is significant difference between in the sales of 4 salesmen.

Conclusion : There is significant difference between the Machine types and no significant difference between the Workers

3. **A laboratory technician measures the breaking strength of each of 5 kinds of linen threads by using 4 different measuring instruments, and obtains the following results, in ounces.**

	I_1	**I_2**	**I_3**	**I_4**
Thread 1	**20.9**	**20.4**	**19.9**	**21.9**
Thread 2	**25**	**26.2**	**27.0**	**24.8**
Thread 3	**25.5**	**23.1**	**21.5**	**24.4**
Thread 4	**24.8**	**21.2**	**23.5**	**25.7**
Thread 5	**19.6**	**21.2**	**22.1**	**22.1**

Perform a 2 – way ANOVA using the 0.05 level of significance for both tests.

Solution:

Null Hypothesis: There is no significant difference between in breaking strength of various threads $H_{01} : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ *and* $H_{02} : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Alternate Hypothesis: There is a significant difference between in breaking strength of various threads $H_{11} : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$ *and* $H_{12} : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

ANOVA TABLE

Source of Variation	**Sum of Square**	**Degrees of Freedom**	**Mean Sum of Squares**	**F – ratio**
Between Rows	66.393	R – 1 = 4	16.598	$F_R = \frac{16.598}{2.078}$ $= 7.987$
Between Columns	5.02	C – 1 = 3	1.673	$F_C = \frac{2.078}{1.673}$ $= 1.242$
Error	24.935	(C-1) (R-1) =12	2.078	
Total	96.348	N – 1 = 11		

Table Value of F = $F_{0.05}(4,12) = 3.26$ and $F_{0.05}(12,3) = 8.74$

Conclusion:

1) $F_R > 3.26$. Hence we reject H_{01} and we conclude that there is significant difference between threads

2) $F_C < 8.74$. Hence we accept H_{02} and we conclude that there is no significant difference between instruments.

4. **Four varieties A,B,C,D of a fertilizer are tested in a randomized block design with 4 replication. The plot yields in pounds are as follows:**

Column / Row	**1**	**2**	**3**	**4**
1	**A(12)**	**D(20)**	**C(16)**	**B(10)**
2	**D(1)**	**A(14)**	**B(11)**	**C(14)**
3	**B(12)**	**C(15)**	**D(19)**	**A(13)**
4	**C(16)**	**B(11)**	**A(15)**	**D(20)**

Analyse the experimental yield.

Solution:

H$_0$: There is no significant difference between the fertilizers and replication

H$_1$:Significant difference between the fertilizers and replication

Variety	**Block**				**Total varieties**				
	1 (X_1)	2 (X_2)	3 (X_3)	4 (X_4)		X_1^2	X_2^2	X_3^2	X_4^2
A	12	14	15	13	54	144	196	225	169
B	12	11	11	10	44	144	121	121	100
C	16	15	16	14	61	256	225	256	196
D	18	20	19	20	77	324	400	361	400
	58	**60**	**61**	**57**	236	868	942	963	865

N=16

T=Grand Total = 236

Correction Factor = $\dfrac{(Grand\ total)^2}{Total\ No\ of\ Observations} = \dfrac{(236)^2}{16} = 3481$

$$TSS = \sum_i \sum_j X_{ij}^2 - C.F = 868 + 942 + 963 + 865 - 3481 = 157$$

$$SSC = \frac{\sum T_{*j}^2}{h} - C.F = 841 + 900 + 930 + 812 - 3481 = 2$$

$$SSR = \frac{\sum T_{i*}^2}{k} - C.F = 729 + 484 + 930.25 + 1482.25 - 3481 = 144.5$$

SSE = TSS – SSC – SSR = 157-2-144.5=10.5

ANOVA Table

Source of Variation	Sum of Squares	Degree of freedom	Mean Square	F- Ratio	F_{Tab}Ratio
Between varieties	SSR=144.5	h - 1= 3	MSR= 48.17	$F_R = 39.27$	$F_{5\%}(3,9) = 3.86$
Between blocks	SSC=2	k – 1=3	MSC = 0.67	$F_C = 0.545$	$F_{5\%}(3, 9) = 3.86$
Residual	SSE = 10.5	(h – 1)(k – 1) = 9	MSE = 1.17		

Conclusion: Cal F_C<Tab F_C and Cal F_R> Tab F_R ⇒ Therefore null hypothesis is rejected. Hence four varieties are not similar. But the varieties are similar along block wise.

Latin Square Design:

Latin Square design controls variation in two direction of the experimental materials as rows and columns resulting in the reduction of experimental error. The analysis of the design results in a three way classification of analysis of variance. Data from Latin Square experiments form a three way classification according the factors rows, columns, and treatments.

CRD	RBD	LSD
To influence one factor	**To influence two factor**	**To influence more than two factor**
No restriction further treatments	**No restriction on treatment and replications**	**The number of replication of each treatment is equal to the number of treatment**
-	**Use only rectangular or Square field**	**Use only Square filed**

The advantages of the Latin square design over other designs are:

(i) With a two way stratification or grouping, the Latin square controls more of the variation than the CRD or the randomized completely block design. The two way elimination of variation often results in small error mean square.

(ii) The analysis is simple.

(iii) Even with missing data the analysis remains relatively simple**.**

Working Procedure (Three – Way classification)

we have seen data from a latin square experiment result in a three way classification result in a three way classification say

(i) variety seeds

(ii) types of spacing(rows)

(iii) the letters for different manure treatment

H_0: There is no difference between columns, between rows and between treatments

H_1 :Not all are equal.

Step 1: Find N= number of observations

Setp 2: Find T = The total value of observations

Step 3: Find the correction Factor = $C.F = \frac{T^2}{N}$

Step 4: Calculate the total sum of squares = $\text{TSS} = \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \ldots\right) - C.\text{F}$

Step 4: Find Total Sum of Square $\text{TSS} = \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \ldots\right) - C.\text{F}$

Step 5: Find column sum of Square $\text{SSC} = \left(\frac{\left(\sum X_1\right)^2}{N_1} + \frac{\left(\sum X_2\right)^2}{N_2} + \frac{\left(\sum X_3\right)^2}{N_3} + \ldots\right) - C.\text{F}$

Where N_i = Total number of observation in each column $(i = 1, 2, 3, \ldots)$

Step 6: Find Row sum of square = $\text{SSR} = \left(\frac{\left(\sum Y_1\right)^2}{N_1} + \frac{\left(\sum Y_2\right)^2}{N_2} + \frac{\left(\sum Y_3\right)^2}{N_3} + \ldots\right) - C.\text{F}$

Where N_j = Total number of observation in each Row $(j = 1, 2, 3, \ldots)$

Step 7: Find SSK for treatments

ANOVA Table for three way classification:

Source of Variation	Sum of Degrees	Degree of freedom	Mean Square	F- Ratio
Column Treatment	**SSC**	**n-1**	$\text{MSC} = \frac{\text{SSC}}{\text{n}-1}$	If MSC>MSE $\text{F}_\text{C} = \frac{\text{MSC}}{\text{MSE}}$ If MSC<MSE $\text{F}_\text{C} = \frac{\text{MSE}}{\text{MSC}}$

Row Treatments	**SSR**	**n-1**	$MSR = \frac{SSR}{n-1}$	If MSR>MSE $F_R = \frac{MSR}{MSE}$ If MSR>MSE $F_R = \frac{MSE}{MSR}$
Between Treatments	**SST**	**n-1**	$MSK = \frac{SSK}{n-1}$	If MSK>MSE $F_K = \frac{MSK}{MSE}$ If MSK>MSE $F_K = \frac{MSE}{MSK}$
Error (or) Residual	**SSE**	**(n-1) (n-2)**	$MSE = \frac{SSE}{(n-1)(n-2)}$	

1. **Analyse the variance in the following latin square of yields (in kgs) of paddy where A, B, C, D denote the different methods of cultivation.**

D 122	**A 121**	**C 123**	**B 122**
B 124	**C 123**	**A 122**	**D 125**
A 120	**B 119**	**D 120**	**C 121**
C 122	**D 123**	**B 121**	**A 122**

Examine whether the different methods of cultivation have given significantly different yields.

Solu.:

H_0: There is no difference between columns, between rows and between treatments

H_1 : Not all are equal.

We shift the origin $X_{ij} = x_{ij} - 100$; n = 4; N = 16

	I	**II**	**III**	**IV**	**Total=T_{i*}**	**$[T_{i*}^2]/n$**	**ΣX_{*ij}^2**
A	2	1	3	2	**8**	16	18
B	4	3	2	5	**14**	49	54
C	0	-1	0	1	**0**	0	2
D	2	3	1	2	**8**	16	18
Total=T_{*j}	**8**	**6**	**6**	**10**	**30**	81	92
$[T_{*j}^2]/n$	16	9	9	25	59		
ΣX_{i*}^2	24	20	14	34	92		

	Letters				**Total=T_{i*}**	**$[T_{i*}^2]/n$**
P	1	2	0	2	**5**	6.25
Q	2	4	-1	1	**6**	9
R	3	3	1	2	**9**	20.25
S	2	5	0	3	**10**	25
Total					**30**	60.5

T=Grand Total = 30 ;Correction Factor = $\frac{(Grand\ \ total)^2}{Total\ \ No\ \ of\ \ Observations} = \frac{(30)^2}{16}$

$$TSS = \sum_i \sum_j X_{ij}^2 - C.F = 92 - \frac{(30)^2}{16} = 35.75$$

$$SSR = \frac{\sum T_{i*}^2}{n} - C.F = 81 - \frac{(30)^2}{16} = 24.75$$

$$SSC = \frac{\sum T_{*j}^2}{n} - C.F = 59 - \frac{(30)^2}{16} = 2.75$$

$$SSL = \frac{\sum T_{i*}^2}{n} - C.F = 60.5 - \frac{(30)^2}{16} = 4.25$$

SSE = TSS – SSC – SSR-SSL = 35.75 – 24.75 – 2.75 – 4.25 = 4

ANOVA Table

Source of Variation	**Sum of Squares**	**Degree of freedom**	**Mean Square**	**F- Ratio**	**F_{Tab}Ratio (5% level)**
Between Rows	SSR=24.75	n - 1= 3	MSR=8.25	F_R= 12.31	F_R(3, 6)=4.76
Between Columns	SSC=2.75	n - 1= 3	MSC = 0.92	F_C = 1.37	

Between Letters	SSL = 4.25	n - 1= 3	MSL = 1.42	F_L = 2.12	Fc(3, 6)=4 .76
Residual	SSE= 4	(n – 1)(n – 2) = 6	MSE = 0.67		F_L(3, 6)=4 .76
Total	35.75				

Conclusion :

Cal F_C< Tab F_C , Cal F_L< Tab F_L and Cal F_R> Tab F_R $\Rightarrow$ There is significant difference between the **rows** , no significant difference between the letters and no significant difference between the columns

2. **The following is a Latin square of a design when 4 varieties of seeds are being tested. Set up the analysis of variance table and state your conclusion. The following is a Latin square of a design when 4 varieties of seeds are being tested. Set up the analysis of variance table and state your conclusion. You may carry out suitable change of origin and scale.**

A 105	**B 95**	**C 125**	**D 115**
C 115	**D 125**	**A 105**	**B 105**
D 115	**C 95**	**B 105**	**A 115**
B 95	**A 135**	**D 95**	**C 115**

(APRIL / MAY '17)

Solu.:

H_0 : Four varieties are similar

H_1 : Four varieties are not similar

Let us take 100 as origin and divide by 5 for simplifying the calculation

Variety	X_1	X_2	X_3	X_4	TOTAL	X_1^2	X_2^2	X_3^2	X_4^2
Y_1	1	-1	5	3	**8**	1	1	25	9
Y_2	3	5	1	1	**10**	9	25	1	1
Y_3	3	-1	1	3	**6**	9	1	1	9
Y_4	-1	7	-1	3	**8**	1	49	1	9
	6	**10**	**6**	**10**	**32**	**20**	**76**	**28**	**28**

N=Total No of Observations = 16 T=Grand Total = 32

$$\text{Correction Factor} = \frac{(\text{Grand total})^2}{\text{Total No of Observatio ns}} = 64$$

$$TSS = \sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2 - C.F = 20 + 76 + 28 + 28 - 64 = 88$$

$$SSC = \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_1} + \frac{(\sum X_3)^2}{N_1} - C.F = \frac{(6)^2}{4} + \frac{(10)^2}{4} + \frac{(6)^2}{4} + \frac{(10)^2}{4} - 64 = 4$$

$$SSR = \frac{(\sum Y_1)^2}{N_1} + \frac{(\sum Y_2)^2}{N_2} + \frac{(\sum Y_3)^2}{N_2} + \frac{(\sum Y_4)^2}{N_2} - C.F = \frac{(8)^2}{4} + \frac{(10)^2}{4} + \frac{(6)^2}{4} + \frac{(8)^2}{4} - 64 = 2$$

To find SSK

Treatment	1	2	3	4	Total
A	1	1	3	7	12
B	-1	1	1	-1	0
C	5	3	-1	3	10
D	3	5	3	-1	10

SSK= $\frac{(\sum Y_1)^2}{K_1} + \frac{(\sum Y_2)^2}{K_2} + \frac{(\sum Y_3)^2}{K_3} + \frac{(\sum Y_4)^2}{K_4} - C.F$

$= 22$

SSE= TSS – SSC–SSR–SSK = 88-4-2-11=60

ANOVA Table

Source of Variation	**Sum of Squares**	**Degree of freedom**	**Mean Square**	**F- Ratio**
Column Treatment	SSC=4	n-1=3	$MSC = \frac{SSC}{n-1}$ =1.33	$F_C = \frac{MSC}{MSE}$ =7.52
Row Treatments	SSR=2	n-1=3	$MSR = \frac{SSR}{n-1}$ =0.67	$F_R = \frac{MSR}{MSE}$ =14.9
Between Treatments	SST=22	n-1=3	$MSK = \frac{SSK}{n-1}$ =7.33	$F_K = \frac{MSK}{MSE}$ =1.36
Error (or) Residual	SSE=60	(n-1) (n-2)=6	$MSE = \frac{SSE}{(n-1)(n-2)}$ $= 10$	

Table value F(3,6) degrees of freedom 8.94

There is significant difference between treatments.

2^2 Factorial Design Experiment:

In factorial experiment, the effect of several factors of variation are investigated simultaneously, the treatment being all the combinations of different factors under study.

Note: 2^2 Factorial = 4 treatment (Let be say 1,a,b,ab)

Procedure:

1. Find N,T

2. Find Correction factor = $C.F = \frac{T^2}{N}$

1. We proceed two way classification between treatment and blocks.

2. For 2×2 *or* 2^2 factorial:

Find

$$SSA = \frac{1}{N}\left[a+ab-b-(1)\right]^2$$

$$SSB = \frac{1}{N}\left[b+ab-a-(1)\right]^2$$

$$SSAB = \frac{1}{N}\left[ab+(1)-a-b\right]^2$$

3. Find $SSE = SST-SSA-SSB-SSAB$

ANOVA Table:

Source of Variation	Sum of Degrees	Degree of freedom	Mean Square	F- Ratio
Between Column	**SSC**	**n-1**	$MSC = \frac{SSC}{n-1}$	If MSC>MSE $F_C = \frac{MSC}{MSE}$ If MSC<MSE $F_C = \frac{MSE}{MSC}$
Between Row	**SSR**	**n-1**	$MSR = \frac{SSR}{n-1}$	If MSR>MSE $F_R = \frac{MSR}{MSE}$ If MSR<MSE

				$F_R = \frac{MSE}{MSR}$
A	**SSA**	**1**	$MSA = SSA$	If MSA>MSE $F_a = \frac{MSA}{MSE}$ If MSA<MSE $F_a = \frac{MSE}{MSA}$
B	**SSB**	**1**	$MSB = SSB$	If MSB>MSE $F_b = \frac{MSA}{MSE}$ If MSB<MSE $F_b = \frac{MSE}{MSA}$
AB	**SSAB**	**1**	$MSAB = SSAB$	If MSAB>MSE $F_{ab} = \frac{MSAB}{MSE}$ If MSAB<MSE $F_{ab} = \frac{MSE}{MSAB}$
Error (or) Residual	**SSE**	**n-c-r+1**	$MSE = \frac{SSE}{n-c-r+1}$	

Problem:

1. **Analyse 2^2 factorial experiment for the following table**

Treatment	**I**	**II**	**III**	**IV**
(l)	**64**	**75**	**76**	**75**
(k)	**25**	**14**	**12**	**33**
(p)	**30**	**50**	**41**	**25**
(kp)	**6**	**33**	**17**	**10**

Solution:

Treatment	I	II	III	IV
(l)	**64**	**75**	**76**	**75**
(k)	**25**	**14**	**12**	**33**
(p)	**30**	**50**	**41**	**25**
(kp)	**6**	**33**	**17**	**10**

We shift the origin $X_{ij} = x_{ij} - 37$;

Treatment	**I**	**II**	**III**	**IV**	**Total=T_{i*}**	**$[T_{i*}^2]/n$**	**ΣX_{*ij}^2**
(l)	27	38	39	38	142	5041	5138
(k)	-12	-23	-25	-4	-64	1024	1314
(p)	7	13	4	-12	12	36	378
(kp)	-31	-4	-20	-27	-82	1681	2106
Total=T_{*j}	-9	24	-2	-5	8	7782	8936
$[T_{*j}^2]/n$	20.25	144	1	6.25	171.5		

T=Grand Total = 8:   N=16

Correction Factor = $\frac{(Grand\ \ total)^2}{Total\ \ No\ of\ \ Observations} = \frac{(8)^2}{16} = 4$

$$TSS = \sum_i \sum_j X_{ij}^2 - C.F = 8936 - 4 = 8932$$

$$SSR = \frac{\sum T_{i*}^2}{n} - C.F = 7782 - 4 = 7778$$

$$SSC = \frac{\sum T_{*j}^2}{n} - C.F = 171.5 - 4 = 167.5$$

SSE = TSS – SSC – SSR = 8932 – 7778 – 167.5 = 986.5

[k] = [kp] – [p] + [k] – [1] =-300   ;[p] = [kp] + [p] - [k] – [1] =-148

[kp] = [kp] – [p] - [k] + [1] =126

$S_k = [k]^2 /4r = 5625$; $S_p = [p]^2 /4r = 1369$; $S_{kp} = [kp]^2 /4r = 992.2$

ANOVA Table

Source of Variation	Sum of Degrees	Degree of freedom	Mean Square	F- Ratio	F_{Tab}
Between Column	**SSC=167.5**	**3**	$MSC = \frac{SSC}{n-1} = 55.83$	If MSC<MSE $F_C = \frac{MSE}{MSC}$ $= 1.963$	$F_{0.05}(3,9)=3.86$
Between Row	**SSR=7778**	**3**	$MSR = \frac{SSR}{n-1}$ $= 2592.67$	If MSR>MSE $F_R = \frac{MSR}{MSE}$ $= 23.65$	$F_{0.05}(3,9)=3.86$
k	**SSk=5625**	**1**	$MSA = SSA = 5625$	If MSA>MSE $F_a = \frac{MSA}{MSE}$ $= 51.32$	$F_{0.05}(1,9)=5.12$
p	**SSp=1369**	**1**	$MSB = SSB = 1369$	If MSB>MSE $F_b = \frac{MSA}{MSE}$ $= 12.49$	$F_{0.05}(1,9)=5.12$
kp	**SSkp=992.2**	**1**	$MSAB = SSAB$ $= 992.2$	If MSAB>MSE $F_{ab} = \frac{MSAB}{MSE}$ $= 9.05$	$F_{0.05}(1,9)=5.12$
Error (or) Residual	**SSE=986.5**	**9**	$MSE = \frac{SSE}{n-c-r+1}$ $= 109.61$		

Conclusion : Cal F_k > Tab F_k , Cal F_p > Tab F_p and Cal F_{kp}> Tab F_{kp}
There is significant difference between the treatments.

UNIT – IV: NON-PARAMETRIC TESTS

PROBLEMS

1. Calculate the expected frequencies for the following data presuming two attributes viz., conditions of home and condition of child as independent.

Condition of Child	Condition of home		
		Clean	Dirty
	Clean	70	50
	Fair	80	20
	Dirty	35	45

Use Chi-Square test at 5% level of significance to state whether the two attributes are independent. (AU NOV/DEC 2013)

Solution:

Null hypothesis: Two attributes are independent.

Alternative hypothesis: Two attributes are not independent.

$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(\frac{(O_{ij}- E_{ij})^2}{E_{ij}}\right)$ follows chi-square distribution with (r-1)(s-1) degrees of freedom.

$\chi^2 = 25.633$

Table value at 5% level of significance is 5.991.

Null hypothesis is rejected.

2. A group of 19 pilots were trained in three different methods video cassette, audio cassette and class room training. Their scores in written exams were as follows:

Video cassette : 74 88 82 93 55 70

Audio cassette : 78 80 65 57 88

Class room training : 68 50 91 84 77 94 81 92

Test whether there is any difference in the effectiveness of the three methods. Use appropriate rank sum test.

Solution:

Here three independent populations are given

$\therefore$ we use Kruscal – wallis H-Test.

Null Hypothesis: All the Populations are identical.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

Alternate Hypothesis: All the Populations are not identical.

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$

Level of significance: $\alpha = 0.05$

The test statistics:

$$H \text{ or } W = \frac{12}{n(n+1)}\left[\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(n+1)$$

Where

$k = 3$ (Number of populations or samples)

R_i = sum of the ranks of all items in sample i

Analysis:

Ranking the data jointly from 1 to 19, we find that

R_1 = sum of the ranks of all items in sample 1

=7+14+12+18+2+6

=59

R_2 = sum of the ranks of all items in sample 2

=9+10+4+3+15

=41

R_3 = sum of the ranks of all items in sample 3

=5+1+16+13+8+19+11+17

=90

n_1= 6 (the no. of items in sample 1)

n_2= 5 (the no. of items in sample 2)

n_3= 8 (the no. of items in sample 3)

$$n = \sum n_i = n_1 + n_2 + n_3 + ... + n_k$$

$$= 6 + 5 + 8$$

$$= 19$$

$$H \ or \ W = \frac{12}{n(n+1)}\left[\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3}\right] - 3(n+1)$$

$$= \frac{12}{19(20)}\left[\frac{(59)^2}{6} + \frac{(41)^2}{5} + \frac{(90)^2}{8}\right] - 3(20)$$

$$= 0.95$$

The sampling distribution of W can be approximated by a χ^2 distribution with $k-1$=3-1=2 degrees freedom. $\alpha = 0.05$

$\therefore \ \chi_\alpha^2 = 5.991$

Conclusion: Here $H < \chi_\alpha^2$ then we Accept Null hypothesis H_0

Hence we conclude that the given three methods are equally effective.

3. Use the sign test to see if there is a 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.

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

Solution:

Null Hypothesis: There is no significant difference between the two types of collections.

(i.e) $H_0 : \ P = 0.5$

Alternate Hypothesis: There is a significant difference between the two types of Collections.

(i.e) $H_1 : P = 0.5$

Level of significance: α =0.05

The test statistics:

Find $d_i = x_i - y_i$

= - - + + + - - - - + + - - + 0

By omitting zero differences, we get n=14

No. of + signs = 6

No. of - signs = 8

p = 0.43   q = 0.57

np=6 and nq =8 is greater than 5

$\therefore$ we use normal distribution

The **standard error** of the proportion $\sigma_p = \sqrt{\frac{pq}{n}}$

$$= \sqrt{\frac{(0.57)(0.47)}{14}}$$

=0.132

$\alpha = 0.05 \quad \therefore \; Z_\alpha = 1.96$

The **confidence interval** $(P - \sigma_p Z_\alpha \, , P + \sigma_p Z_\alpha)$

i.e $(\; 0.5 - (0.132)(1.96) \;\; , 0.5 + (0.132)(1.96) \;)$

i.e $(0.5 - 0.26, \; 0.5 + 0.26)$

i.e $(0.241 \, , \, 0.76)$

Conclusion:

Here the sample proportion p = 0.57 lies within these two limits, so we Accept our Null hypothesis H_0.

Hence there is no significant difference between the two types of collections.

4. Write the merits and demerits of non parametric tests.

Merits and demerits of non parametric tests:

Merits:

1. They do not require us to make the assumption that a population is distributed in the shape of a normal curve or another specific shape.
2. Generally they are easier to do and to understand
3. Sometimes even formal ordering or ranking is not required.

Demerits:

1. They ignore a certain amount of information
2. They are often not us efficient or sharp as parametric tests
3. The non-parametric tests cannot be used to estimate parameters in the populations (or) the confidence intervals for such parameters

It is not possible to solve certain statistical problems by using non parametric tests. A good is the type of problem dealt in the analysis of variance.

5. Two methods of instruction to apprentices are to be evaluated. A director assigns 15 randomly selected trainees to each of the two methods. Due to drop outs, 14 complete in batch 1 and 12 complete in batch 2. An achievement test was given to these successful candidates. Their scores are as follows.

Method 1: 70 90 82 64 86 77 84 79 82 89 73 81 83 66

Method 2: 86 78 90 82 65 87 80 88 95 85 76 94

Test whether the two methods have significant difference in effectiveness. Use Mann-Whitney test at 5% significance level.

Solution:

Null Hypothesis: There is no significant difference between the two methods

(i.e) $H_0 : \mu_1 = \mu_2$

Alternate Hypothesis: There is a significant difference between the two methods

(i.e) $H_0 : \mu_1 \neq \mu_2$

Level of significance: $\alpha = 5\%$

The test statistics:

$$Z = \frac{U - \mu}{\sigma} \quad or \quad z = \frac{U - E(U)}{\sqrt{\mathrm{var}(U)}}$$

Analysis:

Method I	Rank R_1	Method II	Rank R_1
70	4	86	18.5
90	23.5	78	8
82	13	90	23.5
64	1	82	13
86	18.5	65	2
77	7	87	20
84	16	80	10
79	9	88	21
82	13	95	26
89	22	85	17
73	5	76	6
81	11	94	25
83	15		
66	3		
Total	161		190

Here,

$n_1 = 14$ (no. of values in sample I)

$n_2 = 12$ (no. of values in sample II)

$R_1 =$ 161 (Sum of the Ranks of the first sample.)

$R_2 =$ 190 (Sum of the Ranks of the second sample.)

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$= (14)(12) + \frac{14(14+1)}{2} - 161$$

$$= 112$$

$$\mu = \frac{n_1 n_2}{2}$$

$$\mu = \frac{(14)(12)}{2} = \frac{(14)(12)}{2} = 84$$

$$\sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

$$\sigma = \sqrt{\frac{(14)(12)(14+12+1)}{12}} = 19.44$$

Now,

$$Z = \frac{U - \mu}{\sigma} = \frac{112 - 84}{19.44} = 1.44$$

$\alpha = 5\% \quad \therefore Z_\alpha = 1.96$

Conclusion :

Since $|Z| < Z_\alpha$, we Accept Null Hypothesis H_0

Hence, there is no significant difference between the two methods.

6. Kevin Morgan, national sales manager of an electronics firm, has collected the following salary statistics on his field sales force earnings. He has both observed frequencies and expected frequencies if the distribution of salaries is normal. At the 0.05 level of significance, can Kevin conclude that the distribution of sales force earnings is normal?

Earnings in thousands	**25-30**	**31-36**	**37-42**	**43-48**	**49-54**	**55-60**	**61-66**
Observed frequency	**9**	**22**	**25**	**30**	**21**	**12**	**6**
Expected frequency	**6**	**17**	**32**	**35**	**18**	**13**	**4**

Solution:

Null Hypothesis:

H_0 : The distribution of salesforce earnings a normal.

Alternate hypothesis:

H_1: The distribution of salesforce earnings is not normal.

Level of significance: $\alpha = 0.05$

The Test statistic:

$D_n = \max|F_e - F_o|$ with n degrees of freedom

Where F_e = Expected relative frequency

F_o = Observed relative frequency

Analysis:

Observed frequency	Observed cumulative frequency	Observed relative frequency	Expected frequency	Expected cumulative frequency	Expected relative frequency	$D = \|F_e - F_o\|$
9	9	0.072	6	6	0.048	0.024
22	31	0.248	17	23	0.184	0.064
25	56	0.448	32	55	0.440	0.008
30	86	0.688	35	90	0.720	0.032
21	107	0.856	18	108	0.864	0.008
12	119	0.952	13	121	0.968	0.076
6	125	1	4	125	1	0

$\alpha = 0.05$ degrees freedom = n = 7

The table value of D_n = 0.486

$D_n = \max|F_e - F_o| = 0.064$

Conclusion:

Since the calculated value of D_n < table value of D_n

We Accept our Null Hypothesis H_0 .Hence, the distribution of sales force earnings a normal

7. The following contingency table presents the reactions of legislators to a tax plan according to party affiliation. Test whether party affiliation influences the reaction to the tax plan at 0.01 level of signification.

Reaction				
Party	**Infavour**	**Neutral**	**Opposed**	**Total**
Party A	**120**	**20**	**20**	**160**
Party B	**50**	**30**	**60**	**140**
Party C	**50**	**10**	**40**	**100**
Total	**220**	**60**	**120**	**400**

Solution:

Given

Null hypothesis H_0 : Party affiliation and tax plan are independent.

Alternate hypothesis H_1 : Party affiliation and tax plan are not independent.

Level of significance: $\alpha = 0.05$

The test statistics: $\chi^2 = \sum_{i=1}^{r} \sum_{i=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Analysis:

Reaction				
Party	Infavour	Neutral	Opposed	Total
Party A	120	20	20	160
Party B	50	30	60	140
Party C	50	10	40	100
Total	220	60	120	400

E(120)= $\frac{160\times 220}{400} = 88$

E(20)= $\frac{160\times 60}{400} = 24$

E(20)= $\frac{160\times 120}{400} = 48$

E(50)= $\frac{140\times 220}{400} = 77$

E(30)= $\frac{140\times 60}{400} = 21$

E(60)= $\frac{140\times 120}{400} = 42$

E(50)= $\frac{100\times 220}{400} = 55$

E(10)= $\frac{100\times 60}{400} = 15$

E(40)= $\frac{120 \times 100}{400} = 30$

O_{ij}	E_{ij}	O_{ij} - E_{ij}	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
120	88	32	1024	11.64
20	24	-4	16	0.67
20	48	-28	784	16.33
50	77	-27	729	9.47
30	21	9	81	3.86
60	42	18	324	7.71
50	55	-5	25	0.45
10	15	-5	25	1.67
40	30	10	100	3.33

$\therefore \chi^2 = 55.13$

$\alpha = 0.05$ Degrees of freedom = $(r-1)(s-1) = (3-1)(3-1) = 4$

$\therefore \chi^2_\alpha = 13.28$

Conclusion:

Since $\chi^2 > \chi^2_\alpha$, we Reject our Null Hypothesis H_0

Hence, the Party Affiliation and tax plan are dependent.

8. A technician is asked to analyze the results of 22 items made in preparation run. Each item has been measured and compared to engineering specifications. The order of acceptance 'a' and rejections of 'r' is

aarrrarraa aaarrarraa ra

Determine whether it is a random sample or not. Use $\alpha = 0.05$**.**

Solution:

Null Hypothesis H_0 : The Observations are randomly generated.

Alternate hypothesis H_1: The Observations are not randomly generated (two tailed)

Level of significance: $\alpha = 0.05$

The Test statistic:

$$Z = \frac{R-\mu}{\sigma} \quad \text{Where } E(R) = \mu = \frac{2n_1n_2}{n_1+n_2} + 1 \qquad \sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

Analysis:

$$\underset{1}{\underline{aa}}\ \underset{2}{\underline{rrr}}\ \underset{3}{\underline{a}}\ \underset{4}{\underline{rr}}\ \underset{5}{\underline{aaaaa}}\ \underset{6}{\underline{rr}}\ \underset{7}{\underline{a}}\ \underset{8}{\underline{rr}}\ \underset{9}{\underline{aa}}\ \underset{10}{\underline{r}}\ \underset{11}{\underline{a}}$$

R = No. of Runs = 11

$n_1 =$ 12 (no. of items in the first sample)

$n_2 =$ 10 (no. of items in the second sample)

$$E(R) = \mu = \frac{2(12)(10)}{12+10} + 1 = 11.909$$

$$\sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}} = \sqrt{\frac{2(12)(10)(2(12)(10)-12-10)}{(12+10)^2(12+10-1)}} = 2.2688$$

$$Z = \frac{R-\mu}{\sigma} = \frac{11-11.909}{2.2688} = -0.4007 \ |Z| = 0.4007$$

$$\alpha = 0.05 \quad \Rightarrow Z_\alpha = 1.96$$

Conclusion:

Since $|Z| < Z_\alpha$, we Accept our Null hypothesis H_0

Hence, The sample is randomly chosen.

9. From a poll of 800 television viewers, the following data have been accumulated as to, their levels of education and their preference of television stations. We are interested in determining if the selection of a TV station is independent of the level of education

Educational Level				
Public	**High School**	**Bachelor**	**Graduate**	**Total**
Broadcasting	**50**	**150**	**80**	**280**
Commercial Stations	**150**	**250**	**120**	**520**
Total	**200**	**400**	**200**	**800**

(i) State the null and alternative hypotheses.

(ii) Show the contingency table of the expected frequencies.

(iii) Compute the test statistic.

(iv) The null hypothesis is to be tested at 95% confidence. Determine the critical value for this test. (AU JAN 2014)

Solution:

(i)Null Hypothesis: Selection of TV station is independent of level of education

Alternative Hypothesis: Selection of TV station is not independent of level of education

(ii)

Educational Level				
Public	High School	Bachelor	Graduate	Total
Broadcasting	70	140	70	280
Commercial Stations	130	260	130	520
Total	200	400	200	800

(iii)Test statistic = 12.088

(iv) Critical Chi-Square = 5.991, reject Null Hypothesis.

10. The manager of a company believes that differences in sales performance depend upon the salesperson's age. Independent samples of salespeople were taken and their weekly sales record is reported below.

Below 30 years **No. of Sales**	**Between 30 and 45 years** **No. of Sales**	**Over 45 years** **No. of Sales**
24	**23**	**30**
16	**17**	**20**
21	**22**	**23**
15	**25**	**25**
19	**18**	**34**
26	**29**	**36**
	27	**28**

(i) State the null and alternative hypotheses.

(ii) At 95% confidence, test the hypotheses using Kruskal Wallis Test. (AU JAN 2014)

Solution:

(i) Null Hypothesis: All three populations are identical.

Alternative Hypothesis: Not all populations are identical.

(ii) W=6.78, Critical $\chi^2 = 5.991$, at 0.05 level of significance with 2 degrees of freedom.

Hence Reject Null Hypothesis.

UNIT 5 CORRELATION & REGRESSION

PART B

1. Calculate the coefficient of correlation and obtain the lines of regression for the following:

X	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Y	**9**	**8**	**10**	**12**	**11**	**13**	**14**	**16**	**15**

Obtain an estimate of Y which should correspond to the value of X = 6.2

Solution:

X	**Y**	**X²**	**Y²**	**XY**
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
45	**108**	**285**	**1356**	**597**

$$Since\ \ \bar{x} = \frac{\sum x}{n} = \frac{45}{9} = 5 \qquad and \qquad \bar{y} = \frac{\sum y}{n} = \frac{108}{9} = 12$$

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}} = 0.95$$

$$b_{xy} = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum y^2 - \left(\sum y\right)^2} = 0.95 \ \ and \ \ b_{yx} = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} = 0.95$$

$$\text{Re}\,gression\ \ line\ \ y\ on\ x\ is\ \ y - \bar{y} = b_{yx}\left(x - \bar{x}\right) \Rightarrow y = 0.95x + 7.25$$

$$\text{Re}\,gression\ \ line\ \ x\ on\ \ y\ is\ \ x - \bar{x} = b_{xy}\left(y - \bar{y}\right) \Rightarrow x = 0.95y - 6.4$$

$$The\ value\ of\ \ correspond\,ing\ \ to\ \ x = 6.2\ is\ \ y = 0.95 \times 6.2 + 7.25 = 13.14$$

2. Explain the basic components of Time series analysis.

Solution:

The Components of a time series are

1. Secular Trend
2. Seasonal Variations
3. Cyclical Variations
4. Irregular Variations

1. Secular Trend

Trend, also called secular or long term trend, is the basic tendency of a series to grow or decline over a period of time. The concept of trend does not include short range oscillations, but rather the steady movement over a long time. If the values of the time series plotted on a graph paper, cluster around a straight line, the trend is said to be a linear trend. Then the rate of growth in nearly constant. If the plotted points do not fall in the pattern of a straight line, the trend is said to be a non-linear trend or curvilinear trend.

2. Seasonal Variations

Seasonal variations are variations which occur with some degree of regularity within a specific period of one year or shorter. Seasons could be weekly, monthly, quarterly or half yearly depending on the nature of the phenomenon. Production, consumption and prices of commodities show seasonal variations. These variations are periodic and regular.

Eg. Prices of agricultural commodities go down at the time of harvest.

Seasonal variations may occur due to the following causes.

(1) Climate and Weather conditions.
(2) Customs, traditions and habits

A study of seasonal variations is helpful in scheduling purchases, inventory control, personnel recruitment, advertising et. A consumer can gain by purchasing things during slack season.

3. Cyclical Variations

Cyclical variations in a time series are the recurrent variations whose duration is more than one year. A business cycle has four phases – boom, recession, depression and recovery. These phases are uniform but their time duration may vary from cycle to cycle. In spite of the importance of measuring cyclical variations, they are very difficult to measure due to the following reasons.

(i) Business cycles do not show regular periodicity
(ii) The cyclical variations are associated with erratic, random or irregular faces.

4. Irregular Variations

Irregular variations refers to those variations in business or other activities, which do not repeat in a definite pattern. They are caused by random factors like floods, earth quakes, famines, wars, strikes, lockout etc. Sudden changes in demand or very rapid technological progress may also be responsible for these variations. No advance preparation can be done to meet the consequences of irregular variations and their effects are unpredictable and irregular.

3. Given below are the figures of production (in thousand quintals) of a sugar factory.

Year	**1974**	**1975**	**1976**	**1977**	**1978**	**1979**	**1980**
Production	**77**	**88**	**94**	**85**	**91**	**98**	**90**

Fit a straight line by the least squares method and tabulate the trend values.

Solution.

Year	**Production (y)**	**X**	**X^2**	**XY**	**Trend Values**
1974	77	-3	9	-231	83
1975	88	-2	4	-176	85
1976	94	-1	1	-94	87
1977	85	0	0	0	89
1978	91	1	1	91	91
1979	98	2	4	196	93
1980	90	3	9	270	95
Total	**623**	**0**	**28**	**56**	

The straight line trend equation is Y = a + bx

Normal equations are

$$\sum y = na + b\sum x$$
$$\sum xy = a\sum x + b\sum x^2$$, if sum of x is zero then
$$a = \frac{\sum y}{n} = \frac{623}{7} = 89$$
$$b = \frac{\sum xy}{\sum x^2} = \frac{56}{28} = 2$$

Therefore the Trend equation is Y = 89 + 2x

4. Find the two regression lines using the data below: (AU NOV/DEC 2013)

X	**7**	**4**	**8**	**6**	**5**
Y	**6**	**5**	**9**	**8**	**2**

Solution:

$\sum X = 30, \sum Y = 30, \sum X^2 = 190, \sum Y^2 = 210, \sum XY = 192$

$\bar{X} = 6, \bar{Y} = 6$

$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2}$ =0.4

$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$ =1.2 Regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

y=1.2x-1.2

Regression line of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$, x=0.4y-3.6

5. The following data on production (in '000 units) of a commodity from the year 2006-2012. Fit a straight line trend and forecast for the year 2020

Year	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
Production	**6**	**7**	**5**	**4**	**6**	**7**	**5**

(AU NOV/DEC 2013)

Solution:

Trend line Y=a+bX,

Normal equations are $\sum Y = na + b\sum X$

$\sum XY = a\sum X + b\sum X^2$

a=5.71,

b=0.642

Y=5.71+0.642X, X=x-2009

When x = 2020, y=12.772

6. The monthly water consumption in thousand gallons in a hostel for five years is given below. Calculate the seasonal indices by the method of simple averages

Year	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**June**	**July**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
1979	25	23	21	18	15	20	21	25	22	24	32	35
1980	27	25	23	20	17	22	23	27	24	26	35	33
1981	32	31	30	27	25	27	29	30	30	32	41	38
1982	42	40	38	36	34	37	38	40	38	43	52	48
1983	57	50	52	46	49	46	49	55	50	59	64	63

Solution:

Year	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
1979	25	23	21	18	15	20	21	25	22	24	32	35
1980	27	25	23	20	17	22	23	27	24	26	35	33
1981	32	31	30	27	25	27	29	30	30	32	41	38
1982	42	40	38	36	34	37	38	40	38	43	52	48
1983	57	50	52	46	49	46	49	55	50	59	64	63
Total	**183**	**169**	**164**	**157**	**140**	**152**	**160**	**177**	**164**	**184**	**224**	**217**
Avg.	**36.6**	**33.8**	**32.8**	**31.4**	**28**	**30.4**	**32**	**35.4**	**32.8**	**36.8**	**44.8**	**43.4**
Seasonal Index	**105.02**	**96.99**	**94.12**	**90.10**	**80.34**	**87.23**	**91.82**	**101.58**	**94.12**	**105.60**	**128.55**	**124.53**

Seasonal Index = $\frac{Monthly\ \ Average}{General\ \ Average} \times 100$

7. The quarterly sales (in thousands of copies) for a specific education software over the past three years are given in the following table.

	2003	**2004**	**2005**
Quarter 1	**170**	**180**	**190**
Quarter 2	**111**	**96**	**120**
Quarter 3	**270**	**280**	**290**
Quarter 4	**250**	**220**	**223**

(i) Compute the four seasonal factors (Seasonal Indexes). Show all of your computations.

(ii) The trend for these data is Trend = 174+4t (t represents time, where t=1 for Quarter 1 of 2003 And t=12 for Quarter 4 of 2005). Forecast sales for the first quarter of 2006 using the trend and seasonal indexes.

Show all of your computations.(AU JAN 2014).

Solution:

(i)

	2003	2004	2005	Quarter Total	Quarter Average	Seasonal Index
Quarter 1	170	180	190	540	180	0.900
Quarter 2	111	96	120	327	109	0.545
Quarter 3	270	280	290	840	280	1.400
Quarter 4	250	220	223	693	231	1.155

Overall average =200

(ii)Trend = 174+4t=174+4(13) = 226.

Forecast = Trend (SI for Quarter 1) = (226)(0.9)=203.40.

JEPPIAAR
ENGINEERING COLLEGE

Jeppiaar Nagar, Rajiv Gandhi Salai – 600 119

DEPARTMENT OF MASTER OF BUSINESS ADMINSTRATION

QUESTION BANK

I SEMESTER

BA 4101 – STATISTICS FOR MANAGEMENT

Regulation – 2021 (Batch: 2023 -2025)

Academic Year 2023 – 2024

Prepared by

Dr. P. SIVAGAMI, Assistant Professor/MATHEMATICS [S&H]

JEPPIAAR
ENGINEERING COLLEGE

Jeppiaar Nagar, Rajiv Gandhi Salai – 600 119

DEPARTMENT OF MASTER OF BUSINESS ADMINSTRATION

QUESTION BANK

I SEMESTER

BA 4101 – STATISTICS FOR MANAGEMENT

Regulation – 2021 (Batch: 2023 -2025)

Academic Year 2023 – 2024

Prepared by

Dr. P. SIVAGAMI, Assistant Professor/MATHEMATICS [S&H]

UNIT –I : PROBABILITY

PART – A

1. **If *A* and *B* are independent events then prove that A and $\overline{\mathbf{B}}$ are independent.**

Since A and B are independent,

$$P(A\cap B)=P(A)P(B)---(1)$$

$$P(A\cap \bar{B})=P(A)-P(A\cap B) \;\; =P(A)-P(A)P(B) \;\; \text{[using (1)]}$$

$$=P(A)\left[1-P(B)\right]$$

$$P(A\cap \bar{B})=P(A)P(\bar{B}) \;\; \therefore A\,\&\,\bar{B} \text{ are independent events}$$

2. **Let A and B be two events such that $\mathbf{P(A)=\frac{1}{3}, P(B)=\frac{3}{4}, \; P(A\cap B)=\frac{1}{4}}$. Compute $\mathbf{P(A/B)}$ and $\mathbf{P(\bar{A}\cap B)}$.** **(May/June 2019)**

$$P(A/B)=\frac{P(A\cap B)}{P(B)}=\frac{\frac{1}{4}}{\frac{3}{4}}=\frac{1}{3}. \qquad P(\bar{A}\cap B)=P(B)-P(A\cap B)=\frac{3}{4}-\frac{1}{4}=\frac{2}{4}=\frac{1}{2}.$$

3. **If *A* and *B* are two events such that $\mathbf{P(A\cup B)=\frac{3}{4}, \; P(A\cap B)=\frac{1}{4}, P(\bar{A})=\frac{2}{3}}$, find $\mathbf{P(\bar{A}/B)}$.**

$$P(A\cup B)=P(A)+P(B)-P(A\cap B) \Rightarrow \frac{3}{4}=\frac{1}{3}+P(B)-\frac{1}{4} \;\Rightarrow P(B)=\frac{2}{3}$$

$$P(\bar{A}/B)=\frac{P(\bar{A}\cap B)}{P(B)}=\frac{P(B)-P(A\cap B)}{P(B)}=\frac{\frac{2}{3}-\frac{1}{4}}{\frac{2}{3}}=\frac{5}{8}$$

4. **If $\boldsymbol{P(A)=\frac{5}{13}, \;\; P(B)=\frac{3}{7}, P(A\cap B)=\frac{12}{91}}$ find $\boldsymbol{P(A\cup B)}$.**

$P(A\cup B)$= P(A) + P(B) – P(A∩B) = 5/13 +3/7 – 12/91 = 62/91

5. **State Baye's Theorem on Probability.** **(AU NOV/DEC 2013, APR/MAY 2018)**

If $E_1, E_2 \ldots E_n$ are a set of exhaustive and mutually exclusive events associated with a random experiment and A is any other event associated with E_i. Then $P(E_i / A)=\dfrac{P(E_i)P(A/E_i)}{\sum_{i=1}^{n} P(E_i)P(A/E_i)}$, i=1,2,..n

6. **What are mutually exclusive and independent events?** **(AU JAN 2015)**

Two events are said to be mutually exclusive if the occurance of any one of them excludes the occurance of other in a single experiment. Example: Tossing of Coin.

Two (or) more events are independent if the occurance of one does not affect the occurance of the other. Example: If coin is tossed twice; result of second throw is not affected by the result of first throw.

7. **Define Random variable.** **(AU JAN 2014)**

A random variable is a function that assigns a real number X(S) to every element S in the sample space corresponding to a random experiment E.

i.e., X: S → R, S-Sample Space and R-Real Numbers

8. **Explain discrete and continuous variable with examples.** **(AU NOV/DEC 2013)**

A random variable X is said to be discrete, if it takes finite or countable number of values.

Example: X=1,2,3,4,5

A random variable X is said to be continuous, if it takes uncountable number of values.

Example: X is defined in any interval.

9. **Let X be a discrete R.V. with probability mass function**

$$\mathbf{P(X = x) = \begin{cases} \frac{x}{10}, & x = 1,2,3,4 \\ 0, & \text{otherwise} \end{cases}. \text{ Compute } P(X < 3) \text{ and } E\left(\frac{X}{2}\right).}$$

(May/June 2016)

X= x	1	2	3	4
P(X=x)	1/10	2/10	3/10	4/10

$$P(X<3) = P(X=1)+P(X=2) = \frac{1}{10}+\frac{2}{10} = \frac{3}{10}$$

$$E\left(\frac{X}{2}\right) = \frac{1}{2}E(X) = \frac{1}{2}\sum_{x=1}^{4} xP(X=x) = \frac{1}{2}\left\{1\times\frac{1}{10}+2\times\frac{2}{10}+3\times\frac{3}{10}+4\times\frac{4}{10}\right\}$$

$$=\frac{1}{2}\left\{\frac{1+4+9+16}{10}\right\} = \frac{1}{2}\times\frac{30}{10} = \frac{3}{2}.$$

10 **The CDF of a continuous random variable is given by** $\mathbf{F(x) = \begin{cases} 0, & x<0 \\ 1-e^{-\frac{x}{5}}, & x \geq 0 \end{cases}}$

Find the PDF of *X* and mean of *X*.

PDF = $f(x) = \frac{d}{dx}\left[F(x)\right] = \begin{cases} 0, & x<0 \\ \frac{1}{5}e^{-\frac{x}{5}}, & x \geq 0 \end{cases}$

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{0}^{\infty}\frac{1}{5}xe^{-\frac{x}{5}}\,dx = \frac{1}{5}\left[(x)\left(\frac{e^{-\frac{x}{5}}}{-\frac{1}{5}}\right)-(1)\left(\frac{e^{-\frac{x}{5}}}{\frac{1}{25}}\right)\right]_0^{\infty} = \frac{25}{5} = 5$$

11. **The mean of a Binomial distribution is 20 and standard deviation is 4. Find the parameters of the distribution**.

$\therefore \text{np} = 20 \text{ and } \sqrt{\text{npq}} = 4 \Rightarrow \text{npq} = 16 \Rightarrow (20)\text{q} = 16 \Rightarrow \text{q} = \frac{4}{5} \;\; \text{p} = 1-\text{q} = 1-\frac{4}{5} = \frac{1}{5}. \;\; \Rightarrow \text{np} = 20 \Rightarrow \text{n} = 100$

$\therefore$ 100 and $\frac{1}{5}$ are the parameters.

12. **A random variable X has the following probability distribution**

X	:	-2	-1	0	1	2	3
$P\;X$	:	0.1	K	0.2	$2K$	0.3	$3K$

Find K

Since $\sum P(X) = 1$

$$0.1 + K + 0.2 + 2K + 0.3 + 3K = 1$$

$$6K + 0.6 = 1 \Rightarrow 6K = 0.4 \Rightarrow K = \frac{0.4}{6} = \frac{1}{15}$$

13. **Find the probability of getting a total of 5 at least once in three tosses of pair of fair dice .**

$\text{p} = \frac{4}{36} = \frac{1}{9}, \text{q} = \frac{2}{9}, \text{n} = 3$

let X=number of times getting total 5, $X \sim B(n,p)$

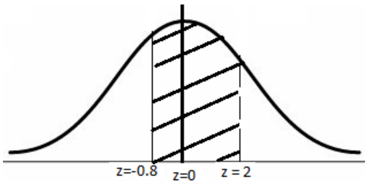
$$\Rightarrow P(X=x) = nC_x p^x q^{n-x}$$

	$P(X\geq 1)=1-P(X<1)=1-P(X=0)=1-3C_0\left(\frac{1}{9}\right)^0\left(\frac{2}{9}\right)^{3-0}=1-\frac{2^3}{9^3}=0.99$
14.	**One percent of jobs arriving at a computer system need to wait until weekends for scheduling, owing to core-size limitations. Find the probability that among a sample of 200 jobs there are no jobs that have to wait until weekends.** $p = 0.01$, $n = 200$, $\lambda = np = 2$, X is the no. of jobs that have to wait $P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}=\frac{e^{-2}(2)^x}{x!}\Rightarrow P(X=0)=\frac{e^{-2}(2)^0}{0!}=e^{-2}=0.1353.$
15.	**The number of monthly breakdown of a computer is having a Poisson distribution with mean equal to 1.8. Find the probability that this computer will function for a month with only one breakdown. (AU MAY/JUNE 2019)** Mean = $\lambda = np = 1.8$, X = No. of breakdowns per month. $P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}=\frac{e^{-1.8}(1.8)^x}{x!}\Rightarrow P(X=1)=\frac{e^{-1.8}(1.8)^1}{1!}=0.2975.$
16.	**If X is a Uniformly distributed random variable with mean 1 and variance $\frac{4}{3}$, find P(X<0).** Mean = $\frac{a+b}{2}=1\Rightarrow a+b=2$ and variance = $\frac{(b-a)^2}{12}=\frac{4}{3}\Rightarrow b-a=4$ By solving the above eqns. We get a = -1 and b = 3 $f(x)=\frac{1}{b-a}\ in\ a<x<b$ $f(x)=\frac{1}{4}in-1<x<3\ \Rightarrow P(X<0)=\int_{-1}^{0}f(x)dx=\int_{-1}^{0}\frac{1}{4}dx=\frac{1}{4}[x]_{-1}^{0}=\frac{1}{4}$
17.	**If R.V ‘X’ is uniformly distributed over (-3,3), then compute P ($\lvert X-2\rvert<2$).** $f(x)=\frac{1}{b-a}\ in\ a<x<b=\frac{1}{6}\ in\ -3<x<3$ $P(\lvert X-2\rvert<2)=P(-2<X-2<2)=P(0<X<4)=\int_0^3 f(x)dx=\int_0^3\frac{1}{6}dx=\frac{1}{6}[x]_0^3=\frac{1}{2}.$
18.	**Define Normal distribution.** A random variable X is said to have a Normal distribution with parameters μ (mean) and σ^2 (variance) if its probability density function is given by the probability law $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},-\infty<x<\infty,-\infty<\mu<\infty,\sigma>0$
19.	**State Poisson distribution as limiting form of binomial distribution.** Poisson distribution is a limiting case of Binomial distribution under the following conditions: (i). n the number of trials is indefinitely large, (i.e.) $n\to\infty$ (ii). P the constant probability of success in each trial is very small (i.e.) P→0 (iii). $np=\lambda$ is finite.
20.	**X is a normal variate with mean = 30 and S.D = 5. Find $P[26\leq X\leq 40]$** X follows N(30, 5) $\therefore \mu=30$ & $\sigma=5$ Let $Z=\frac{X-\mu}{\sigma}$ be the standard normal variate

$$P[26 \le X \le 40] = P\left[\frac{26-30}{5} \le Z \le \frac{40-30}{5}\right] = P[-0.8 \le Z \le 2]$$

$$= P[-0.8 \le Z \le 0] + P[0 \le Z \le 2]$$

$$= P[0 \le Z \le 0.8] + [0 \le Z \le 2] = 0.2881 + 0.4772 = 0.7653.$$



PART B

1a. **A bag contains 3 black and 4 white balls. Two balls are drawn at random one at a time without replacement. (1) What is the probability that the second ball drawn is white? (2) What is the conditional probability that the first ball drawn is white if the second ball is known to be white? (May/June 2019)**

1b. **An industrial unit has 3 machines – 1, 2 and 3, which produce the same item. It is known that machines 1 and 2 each produce 30% of the total output, while machine 3 produces 40% of the remaining output. It is also known that 2% of machine 1 output is defective, while machines 2 and 3, each produce 3% defective items. All the items are put into one stockpile and then one item is chosen at random. Find the probability that defective item is produced by machine 1 (AU JAN 2016)**

1c. **The first bag contains 3 white balls, 2 red balls and 4 black balls. Second bag contains 2 white, 3 red and 5 black balls and third bag contains 3 white, 4 red and 2 black balls. One bag is chosen at random and from it 3 balls are drawn. Out of three balls two balls are white and one is red. What are the probabilities that they were taken from first bag, second bag, third bag.**

2a. **A consulting firm rents cars from three rental agencies in the following manner: 20% from agency D, 20% from agency E and 60% from agency F. If 10% cars from D, 12% of the cars from E and 4% of the cars from F have bad tyres, what is the probability that the firm will get a car with bad tyres? Find the probability that a car with bad tyres is rented from agency F.**

2b. **A random variable X has the following probability function:**

X	:	0	1	2	3	4	5	6	7
P X	:	0	K	2K	2K	3K	K^2	$2K^2$	$7K^2+K$

Find (i) K , (ii) Evaluate $P(X<6), P(X \ge 6)$ and $P(0<X<5)$ (iii) Determine the distribution function of X. (iv) $P(1.5<X<4.5/X>2)$ (v) $E(3X-4)$, Var(3X - 4)

(vi) If $P[X \le C] > \frac{1}{2}$, find the minimum value of *C*. (April/May 2015)

3a. **A random variable *X* has the probability mass function $f(x) = \frac{1}{2^x}$, x= 1,2,3,...**

Find its (i) M.G.F (ii) Mean (iii) Variance.

3b. **If the density function of a continuous random variable X is given by** $f(x) = \begin{cases} ax, & 0 \le x \le 1 \\ a, & 1 \le x \le 2 \\ 3a - ax, & 2 \le x \le 3 \\ 0, & \text{elsewhere} \end{cases}$

Find the value of a and find the c.d.f of X , *P(X ≤ 1.5)*.

4a. **In the production of electric bulbs , the quality specification of their life was found to normally distributed with average life of 2100 hours and standard deviation of 80 hours . In a sample of 1500 bulbs, find out the expected number of bulbs likely to burn for**

(i) more than 2200 hours ,

(ii) less than 1950 hours ,

(iii) more than 2000 hours but less than 2150 hours . (AU MAY/JUNE 2019) (APR /MAY 2018)

4b. **Trains arrive at a station at 15 minutes interval starting at 4 a.m. If a passenger arrive at a station at a time that is uniformly distributed between 9.00 a.m. and 9.30 a.m., find the probability that he has to wait for the train for (i) less than 6 minutes (ii) more than 10 minutes.**

5a. **In a normal population with mean 15 and standard deviation 3.5, it is found that 647 observations exceed 16.25. What is the total number of observations in the population?**

5b. **Derive MGF, Mean and Variance of Binomial and Poison Distribution.**

UNIT –II : SAMPLING DISTRIBUTION AND ESTIMATION

PART – A

1. **Define Population.** **(AU JAN 2016)**
The group of individuals under study is called population. The population may be finite or infinite.

2. **Define Sample and Sample Size.** **(AU JAN 2018)**
A finite subset of statistical individuals in a population is called **Sample**. The number of individuals in a sample is called **Sample Size (n)**.

3. **Define Parameter and Statistic.** **(AU JAN 2016)**
A numerical measure of a population is called a population parameter or simply a parameter.
A numerical measure of the sample is called a sample statistic or simply a statistic.

4. **Define Sampling distribution.**
The sampling distribution of a statistic is the probability distribution of all possible values the statistic may take, when computed from random samples of same size, drawn from a specified population. Like any other distribution, a sampling distribution will have its mean, standard deviation and moments of higher order.

5. **Distinguish between estimate and estimator**.
An estimator of a population parameter is a sample statistic used to estimate the parameter. An estimate of the parameter is a particular numerical value of the estimator obtained by sampling**.**

6. **What is Standard Error and mention its uses?** **(AU MAY/JUNE 2019)**
The standard deviation of the sampling distribution of a statistic is known as its standard error.
The magnitude of the standard error gives an index of the reliability of the estimate of the parameter. The greater the standard error of the estimate, lesser will be the reliability of the sample.Standard error is useful for determining the probable limits or confidence limits for an unknown parameter with a specified confidence co-efficient. Standard error is also used for testing of hypothesis.

7. **Define Type I error and Type II error.** **(AU MAY/JUNE 2019)**
Type I error: If we reject a hypothesis when it should be accepted, we say that type I error.
Type II error: If we accept a hypothesis when it should be rejected, we say that a type II error.

8. **Define Critical region.**
A region corresponding to a test statistic in the sample space which tends to rejection of H_0 (Null Hypothesis) is called critical region or region of rejection.
The region complementary to the critical region is called the region of acceptance.

9. **Define level of significance.**
The probability ‘α’ (the probability of making type I error) that a random value of the test statistic belongs to the critical region is known as the level of significance. In other words, level of significance is the size of the type I error.
The levels of significance usually employed in testing of hypothesis are 5% and 1%.

10. **Define Critical values or significant values.**
The value of test statistic, which divides the critical (or rejection) region and acceptance region, is called the critical value or significant value. It depends on the level of significance used and the alternative hypothesis.

11. **Write the two properties of the sampling distribution of the mean when the population is normally distributed.** **(AU JAN 2016)**

1. It has a mean equal to the population mean $\mu_{\bar{x}} = \mu$.

2. It has a standard deviation equal to the population standard deviation divided by the square root of the sample size $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where $\sigma_{\bar{x}}$ is the standard error of the mean.

12. **State the characteristics of best estimator.**

i)Unbiasedness ii)Efficiency iii)Consistency iv)Sufficiency

13. **Define One tailed test and two tailed test.** **(AU NOV/DEC 2013)**

When the hypothesis about the population parameter is rejected only for the value of sample statistic falling into one of the tails of the sampling distribution, then it is known as one-tailed test.

If it is right tail then it is called right-tailed test or one-sided alternative to the right and if it is on the left tail, then it is one-sided alternative to the left and called left-tailed test. Two tailed test is one where the hypothesis about the population parameter is rejected for the value of sample statistic falling into the either tails of the sampling distribution.

14. **Given that $n_1 = 400,\ \overline{x_1} = 250,\ s_1 = 40$ for one sample and $n_2 = 400,\ \overline{x_2} = 220,\ s_2 = 55$ for another sample, find the standard error of $\bar{x}_1 - \bar{x}_2$.**

The standard error of $\bar{x}_1 - \bar{x}_2$ **is**

$$S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ \text{, where } S = \sqrt{\frac{s_1{}^2+s_2{}^2}{n_1+n_2-2}} = 2.4074$$

$$\text{S.E of } \left(\bar{x}_1 - \bar{x}_2\right) = 2.4074\sqrt{\frac{1}{400}+\frac{1}{400}} = 0.17023$$

Therefore standard error of $\bar{x}_1 - \bar{x}_2$ is 0.17023.

15. **Write 95% confidence interval of the population mean.** **(AU MAY/JUNE 2014)**

$$\bar{x} - t_{0.05}\frac{S}{\sqrt{n-1}} \le \mu \le \bar{x} + t_{0.05}\frac{S}{\sqrt{n-1}}$$

16. **You want to determine whether the mean of the population from which this sample was taken is significantly different from 48. State the null and the alternate hypothesis.** **(AU MAY/JUNE 2014)**

Null hypothesis $H_0: \mu = 48$ and Alternate hypothesis $H_1: \mu \neq 48$.

17. **For test market find the sample size needed to estimate the true proportion of consumers satisfied with a certain new product within ± 0.04 at 90% confidence level.**

If proportion is not given, take p = q = 0.5. E = 0.04, Z_α = 1.645

$$n = \frac{Z_\alpha{}^2.pq}{E^2} = \frac{(1.645)^2(0.5)(0.5)}{(0.04)^2} = 423$$

18. **State central limit theorem**. **(AU JAN 2014, 2015)**

A sample of samples is always normally distributed about the mean of sample means, even if the samples themselves are not normally distributed themselves about their means.

19. **Differentiate between point estimate and interval estimate** **(AU JAN 2015)**

Point estimate	**Interval estimate**
When a single value is used as an estimate, the estimate is called a point estimate of the population parameter. For example, the sample mean is the sample statistic used as an estimate of population mean	An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an interval estimate of the parameter.

20	**Write the confidence interval for the population mean for large samples when σ is known.** The confidence interval for μ when σ is known and sampling is done from a normal population or with a large sample, is $\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ .Here $\bar{x}$ - sample mean, σ - standard deviation, n – size of the sample.
	PART – B
1a.	**Below you are given the values obtained from a random sample of 4 observations taken from an infinite population.** **32, 34, 35, 39** **(i) Find a point estimate for μ. Is this an unbiased estimate of μ? Explain.** **(ii) Find a point estimate for σ^2. Is this an unbiased estimate of σ^2? Explain.** **(iii) Find a point estimate for σ.** **(iv) What can be said about the sampling distribution of $\bar{x}$? Be sure to discuss the expected value, the standard deviation and the shape of the sampling distribution of $\bar{x}$. (AU JAN 2014)**
1b.	**In a sample of 25 observations from a normal distribution with mean 98.6 and standard deviation 17.2. What is $[92 < \bar{x} < 102]$? (AU JAN 2016)**
2a.	**Explain the types of estimation and the qualities of a good estimator. (AU NOV/DEC 2013)**
2b.	**In a random sample of 75 axle shafts, 12 have a surface finish that is rougher than the specifications will allow. Suppose that a modification is made in the surface finishing process and subsequently a second random sample of 85 axle shafts is obtained. The number of defective shafts in this second sample is 10. Obtain an approximate 95% confidence interval on the difference in the proportions of defectives produced under the two processes. (AU JAN 2016)**
3a.	**In a batch chemical process used for etching printed circuit boards, two different catalysts are being compared to determine whether they require different emersion times for removal of identical quantities of photoresist material. Twelve batch were run with catalyst 1, resulting in a sample mean emersion time of 24.6 minutes and sample standard deviation of 0.85 minutes. Fifteen batches were run with catalyst 2, resulting in a mean emersion time of 22.1 minutes and a standard deviation of 0.98 minutes. Find a 95% confidence interval on the difference in means, assuming that $\sigma_1{}^2 = \sigma_2{}^2$. Also find a 90% confidence interval on the ratio of variances. (AU JAN 2016)**
3b.	**From a population of 540, a sample of 60 individuals are taken. From this sample, the mean is found to be 6.2 and the standard deviation is 1.368.** **(1) Find the estimated standard error of the mean.** **(2) Construct a 90 percent confidence interval for the mean. (AU MAY/JUNE 2019)**
4.	**Discuss various non – probability sampling methods in use with its applications.**
5a.	**A bank has kept records of the checking balances of its customers and determined that the average daily balance of its customers is Rs. 300 with a standard deviation of Rs. 48. A random sample of 144 checking accounts is selected.** **(i) What is the probability that the sample mean will be more than Rs. 306.60?** **(ii) What is the probability that the sample mean will be less than Rs. 308?** **(iii) What is the probability that the sample mean will be between Rs.302 and Rs. 308?** **(iv) What is the probability that the sample mean will be atleast Rs. 296? (AU MAY/JUNE2014)**
5b.	**A certain city is studied for demographic characteristics. It is found that the age has a standard deviation of 5.3 years and 60% of the population is female. What should be the sample size if the age is to be estimated with an error of less than 1 year? What should be the sample size if a similar estimation is to be done on the proportion of female population if the desired accuracy is to be within 5%?If the sample average age is found to be 37.25 for a sample size of 300, estimate the population age range with a confidence level of 95%. (AU MAY 2020)**

5c.	**The life time of a certain brand of an electric bulb may be considered as a random variable with mean 1200h and standard deviation 250h. Find the probability, using central limit theorem, that the average lifetime of 60 bulbs exceeds 1250hours.**
	UNIT – III: TESTING OF HYPOTHESIS – PARAMETRIC TESTS.
	PART – A
1.	**Define t-statistic.** The t – distribution is used when sample size is 30 or less and the population standard deviation is unknown. The t – statistic is defined as $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ where $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$. The t – distribution has been derived mathematically under the assumption of a normally distributed population.
2.	**List out the applications of t –distribution. (AU NOV/DEC 2017) (AU APR/MAY 2018)** 1. To test the significant difference between the means of two independent samples. 2.To test the significant difference between the means of two dependent samples or paired observation. 3. To test the significance of the mean of a random sample. 4 To test the significance of an observed correlation coefficient.
3.	**Mention the Properties of t – distribution.** 1. The t distribution ranges from $-\infty$ to ∞ 2. The t – distribution like the standard normal distribution is bell shaped, symmetrical around mean zero. 3. The variance of the t – distribution is greater than one and is defined only when $v \geq 3$
4.	**What is the purpose of F – test?** F test refers to a test of hypothesis concerning two variances derived from two samples. It is used to test whether the two sample variances are equal or not that is $F = \frac{S_1^{\,2}}{S_2^{\,2}}$, $S_1 > S_2$. Thus F statistics is the ratio of independent estimates of population variances.
5.	**What are the assumptions on which F-test is based?** 1. Normality: The values in each group should be normally distributed. 2. Independence of error: The variations of each value around its own group mean. i.e. error should be independent of each value. 3. Homogeneity**:** The variances within each group should be equal for all groups.
6.	**When to use the normal and 't' distribution in making tests of hypothesis about means? (JAN 2016)** When the sample size is greater than 30 we say it is large sample and when it is less than we say it as small sample. For large samples we use normal distribution and for small samples we use t test.
7.	**Estimate the standard error of the difference between the two proportions if $\overline{p_1} = 0.10$, $\overline{p_2} = 0.1333$,$n_1 = 50$ and $n_2 = 75$? (AU JAN 2016)** Let us first calculate the weighted average of p_1 and p_2 that is say $p = \frac{n_1p_1+n_2p_2}{n_1+n_2}$ =0.1199 S.E. is $\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ = 0.05931.
8.	**What do you mean by degrees of freedom?** Degrees of freedom are the total number of observations minus the number of independent constraints imposed on the observations.
9.	**Define Experimental Error** The estimation of the amount of variations due to each of the independent factors separately and then comparing these estimates due to assignable factors with the estimate due to chance factor is known as experimental error.

10. **State the assumptions of Student's t – test.**

    1. The sample observations are independent.
    2. The parent population from which the sample is drawn is normal.
    3. The population standard deviation σ is unknown.

11. **Define Local Control.**

When the number of treatments becomes large, it may not be possible to accommodate all the treatments in one block because that will increase heterogeneity within blocks. The process of making the experimental units homogeneous and reducing the experimental error is known as local control.

12. **What are the Properties of F- distribution? (AU MAY/JUNE 2019)**

1. The value of F must always be positive or zero since variances are squares and can never assume negative values. Its value will always lie between 0 and ∞.
2. The shape of the F- distribution depends upon the number of degrees of freedom.
3. The F – distribution is positively skewed.

13. **Define Analysis of Variance**.

Analysis of Variance is a technique that will enable us to test for the significance of the difference among more than two sample means.

14. **What are the assumptions of analysis of variance?**

(i) The sample observations are independent

(ii) The Environmental effects are additive in nature

(iii) Sample observation are coming from normal

15. **Distinguish between z-test and t-test. (AU JAN 2015)**

z-test	**t-test**
(i) Used for large samples	(i) Used for small samples
(ii) Follows normal distribution.	(ii)Follows student's t distribution.

16. **Define one way classification and two way classifications in ANOVA.**

The entire experiment influences on only single factor is one way classification. The entire experiment influences on only two factors is two way Classification.

17. **What are the basic principles of design of experiments? (APR/MAY 2018)**

(i) Randomization (ii) Replication (iii) Local Control

18. **What are the usual assumptions made in the analysis of a randomized block Experiment?**

(i) All the experimental units are homogenous

(ii) Each treatment replicates ' r ' times.

19. **Write down the ANOVA table for One way classification**

Source of Variation	Sum of Degrees	Degree of freedom	Mean Square	F- Ratio
Between Samples	SSC	K-1	$MSC = \frac{SSC}{K-1}$	$F_C = \frac{MSC}{MSE}$
Within Samples	SSE	N-K	$MSE = \frac{SSE}{N-K}$	

20. **Write down the ANOVA table for Randomized Block Design**

Source of Variation	Sum of Degrees	Degree of freedom	Mean Square	F- Ratio
Column Treatment	SSC	c-1	$\text{MSC} = \frac{\text{SSC}}{c-1}$	$F_C = \frac{\text{MSC}}{\text{MSE}}$
Row Treatments	SSR	r-1	$\text{MSC} = \frac{\text{SSR}}{r-1}$	$F_R = \frac{\text{MSR}}{\text{MSE}}$
Error (or) Residual	SSE	(r-1) (c-1)	$\text{MSE} = \frac{\text{SSE}}{(r-1)(c-1)}$	

PART – B

1a. **The following are the number of mistakes made in 5 successive days by 4 technicians working for a photographic laboratory. Test whether the difference among the four sample means can be attributed to chance. (Test at a level of significance $\alpha = 0.01$)**

Technicians			
I	**II**	**III**	**IV**
6	**14**	**10**	**9**
14	**9**	**12**	**12**
10	**12**	**7**	**8**
8	**10**	**15**	**10**
11	**14**	**11**	**11**

1b. **40 people were attacked by a disease and only 36 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease, is 85%at 5% level of significance?**

2a. **Two independent samples of 8 and 7 items respectively had the following values.**

Sample I	**9**	**11**	**13**	**11**	**15**	**9**	**12**	**14**
Sample II	**10**	**12**	**10**	**14**	**9**	**8**	**10**	**-----**

Is the difference between the means of samples significant? (AU APR / MAY 2018)

2b. **The following table shows the yields per acre of four different plant crops grown on lots treated with three different types of fertilizer. Determine at the 5% significance level whether there is a difference in yield per acre**

(i) due to the fertilizers and

(ii) due to the crops

	Crop - I	**Crop - II**	**Crop - III**	**Crop - IV**
Fertilizer A	**4.5**	**6.4**	**7.2**	**6.7**
Fertilizer B	**8.8**	**7.8**	**9.6**	**7.0**
Fertilizer C	**5.9**	**6.8**	**5.7**	**5.2**

3a. **Given a sample mean of 83, a sample standard deviation of 12.5 and a sample size of 22, test the hypothesis that the value of the population mean is 70 against alternative that it is more than 70. Use the 0.025 significance level. (AU JAN 2016)**

3b. **A random sample of 10 boys had the following I.Q's: 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q of 100? Find the reasonable range in which most of the mean I.Q values of samples of 10 boys lie.**

4a. **The following table gives biological values of a protein from cow's milk and buffalo's milk at certain level . Examine if the average values of protein in the 2 samples significantly differ. (APR /MAY 2018)**

Cow's milk	**1.82**	**2.02**	**1.88**	**1.61**	**1.81**	**1.54**
Buffalo's milk	**2.00**	**1.83**	**1.86**	**2.03**	**2.19**	**1.88**

4b. **A lathe is set to cut bars of steel into lengths of 6 centimeters. The lathe is considered to be in perfect adjustment if the average length of the bars it cuts is 6 centimeters. A sample of 121 bars is selected randomly and measured. It is determined that the average length of the bars in the sample is 6.08 centimeters with a standard deviation of 0.44 centimeters.**

(i) Formulate the hypotheses to determine whether or not the lathe is in perfect adjustment.

(ii) Compute the test statistic. (iii) What is your conclusion? (AU JAN 2014)

5a. **The daily production rates for a sample of factory workers before and after a training program are shown below. Let d=After – Before.**

Worker	**Before**	**After**
1	**6**	**9**
2	**10**	**12**
3	**9**	**10**
4	**8**	**11**
5	**7**	**9**

We want to determine if the training program was effective.

(i) Give the hypotheses for this problem.

(ii) Compute the test statistic.

(iii) At 95 % confidence, test the hypotheses. That is, did the training program actually increase the production rates? (AU JAN 2019)

5b. **The following table shows the lifetimes in hours of samples from three different types of television tables manufactured by a company. Determine whether there is a difference between the three types at significance level of 0.01**

Sample 1	**407**	**411**	**409**		
Sample 2	**404**	**406**	**408**	**405**	**402**
Sample 3	**410**	**408**	**406**	**408**	

UNIT – IV: NON-PARAMETRIC TESTS

PART – A

1. **State any two properties of χ^2 distribution.**

1. The exact shape of the distribution depends upon the number of degrees of freedom n. In general when n is small, the shape of the curve is skewed to the right and as n gets larger, the distribution becomes more and more symmetrical.

2. The mean and variance of the χ^2 distribution are n and 2n respectively.

3. The sum of the independent χ^2variates is also a χ^2variate.

2. **Explain the various uses of Chi-square test.**

1.Test of goodness of fit

2.Test of independence of attributes

3. Test of Homogeneity of independent estimates of the population correlation coefficient.

3. **What are the conditions for the validity of Chi-square test? (AU MAY/ JUNE 2016)**

1. The experimental data must be independent of each other.

2. The total frequency must be reasonably large, say $\geq$ 50.

3. No individual frequencies should be less than 5, If any frequency is less than 5, then it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5.Finally adjust for the degrees of freedom lost in pooling.

4. **Write the formula for chi square test of single standard deviation.** **(AU MAY/JUNE 2014)**

The formula is $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

5. **What are the uses of χ^2 test?**

1. To test the homogeneity of independent estimates of the population variances.
2. To test the goodness of fit.
3. To test for independence of attributes.

6. **Explain the Chi – square test as a test of independence.**

It is applied to test the association between the attributes when the sample data is presented in the form of a contingency table with any number of rows or columns.

$E_{ij} = \frac{R_i \times C_j}{G.T}$ where R_i = i^th row total and C_j = j^th column total , GT = Grand Total

$\chi^2_{\text{calculated value}} < \chi^2_{\text{tabulated value}}$

then accept H_0.

7. **What are the disadvantages of Non Parametric tests?**

1. They ignore a certain amount of information
2. They are often not as efficient or sharp as parametric tests.
3. The non-parametric tests cannot be used to estimate parameters in the population or the confidence intervals for such parameters.

8. **What is meant by Non Parametric test?** **(AU NOV/DEC 2013)**

Non parametric test is the test that does not make any assumption regarding from which the sampling is done. They are often called as distribution -free methods.

9. **Name any four Non parametric test.** **(AU JAN 2015)**

1. Sign Test for paired data
2. Rank Sum tests
   (a)Mann-Whitney U-Test (b)Kruskal -Wallis Test or H test.
3. Rank correlation test
4. One sample run test.

10. **Define the statistics used in the U – test and give its mean.**

U = $n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$, Mean = $\frac{n_1 n_2}{2}$

11. **Define the statistics used in the H – test.** **(AU MAY/JUNE 2019)**

H= $\frac{12}{n(n+1))}\left[\sum_{i=1}^{k}\frac{R_i^{\,2}}{n_i}\right] - 3(n+1)$

12. **When Kruskal –Wallis test is used ?** **(NOV/DEC 2017)**

The Kruskal –Wallis test is used to test whether the 3 or more populations are identical or not . The K-W test is based on the analysis of independent random samples from each of the k populations.

13. **Define Kolmogorov smirnov Test.**

It is a simple non parametric test for testing whether there is a significance between an observed frequency distributions and a theoretical frequency distribution. It is another measure of the goodness of fit. (i.e.,) $D_n = max|F_e - F_o|$

14. **When Mann-Whitney U-Test is used ?**

The U test is used to test whether the 2 populations are identical or not . The U test is based on the analysis of independent random samples from two polpulations.

15. **What are the advantages of Kolmogorov-smirnov test ?**

(i) It is a more powerful test.

(ii) It is easier to use since it does not require that the data be grouped in any way.

16. **Write any two advantages of non-parametric methods over parametric methods.**

1. They do not require us to make the assumption that a population is distributed in the shape of a normal curve or another specific shape.
2. Generally they are easier to do and to understand.
3. Sometimes even formal ordering or ranking is not required.

17. **When sign test is used?**

1. When there are pair of observations on two things being compared.
2. For any given pair, each of two observations is made under similar conditions.
3. No assumptions are made regarding the parent population.

18. **Write the formula for run test .** **(AU APR/MAY 2018)**

Let R be the number of runs , n_1 = number of items in first sample , n_2 = number of items in second sample .

Here ,R is approximated by normal distribution

$$Z=\frac{R-E(R)}{\sqrt{V(R)}}\approx N(0,1)$$

where $E(R)=\mu=\frac{2n_1n_2}{n_1+n_2}+1, \quad V(R)=\sigma^2=\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$

$$\therefore Z=\frac{R-\mu}{\sigma}\approx N(0,1)$$

19. **Find the number of runs for the following series MMFFFMFFMMMM.**

Number of runs = R = 5.

20. **Find the number of runs for the following series HTHHHHHTTTHTHTHHTTTHHTTTH .**

Number of runs = R = 13 .

PART – B

1.a **Two sample polls of votes for two candidates A and B for a public office were taken, one from among the residents of rural areas and one from the residents of urban areas . The results are given in the table . Examine whether the nature of the area is related to voting preference in this election.** **(AU NOV /DEC 2017)**

Votes for area	A	B	TOTAL
Rural	620	380	1000
Urban	550	450	1000
TOTAL	1170	830	2000

1b. **An experiment designed to compare three preventive methods against corrosion yielded the following maximum depths of pits (in thousands of an inch) in pieces of wire subjected to the respective treatments:**

Method A:	77	54	67	74	71	66	
Method B:	60	41	59	65	62	64	52
Method C:	49	52	69	47	56		

Use the kruskal – Wallis test at the 5% level of significance to test the null hypothesis that the three samples come from identical populations.

2a. **Use the sign test to see if there is a 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.**

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

2b. **Test whether the following numbers 0.44, 0.81, 0.14, 0.05, 0.93 are uniformly distributed using Kolmogorov – smirnov test** **(AU NOV – DEC 2018)**

3a. **Two methods of instruction to apprentices are to be evaluated. A director assigns 15 randomly selected trainees to each of the two methods. Due to drop outs, 14 complete in batch 1 and 12 complete in batch 2. An achievement test was given to these successful candidates. Their scores are as follows. Method 1: 70 90 82 64 86 77 84 79 82 89 73 81 83 66**

Method 2: 86 78 90 82 65 87 80 88 95 85 76 94

Test whether the two methods have significant difference in effectiveness. Use Mann-Whitney test at 5% significance level.

3b. **Kevin Morgan, national sales manager of an electronics firm, has collected the following salary statistics on his field sales force earnings. He has both observed frequencies and expected frequencies if the distribution of salaries is normal. At the 0.05 level of significance, can Kevin conclude that the distribution of sales force earnings is normal? (AU MAY/JUNE 2019)**

Earnings in thousands	25-30	31-36	37-42	43-48	49-54	55-60	61-66
Observed frequency	9	22	25	30	21	12	6
Expected frequency	6	17	32	35	18	13	4

4a. **The following contingency table presents the reactions of legislators to a tax plan according to party affiliation. Test whether party affiliation influences the reaction to the tax plan at 0.01 level of signification.**

	Reaction			
Party	Infavour	Neutral	Opposed	Total
Party A	120	20	20	160
Party B	50	30	60	140
Party C	50	10	40	100
Total	220	60	120	400

4b. **A technician is asked to analyze the results of 22 items made in preparation run. Each item has been measured and compared to engineering specifications. The order of acceptance ‘a’ and rejections of ‘r’ is** $aarrrarraaaarrarraara$ **Determine whether it is a random sample or not. Use** $\alpha = 0.05$**.**

5a. **From a poll of 800 television viewers, the following data have been accumulated as to, their levels of education and their preference of television stations. We are interested in determining if the selection of a TV station is independent of the level of education (AU JAN 2016)**

	Educational Level			
	High School	Bachelor	Graduate	Total
Public Broadcasting	50	150	80	280
Commercial Stations	150	250	120	520
Total	200	400	200	800

(i) State the null and alternative hypotheses.

(ii) Show the contingency table of the expected frequencies. (iii) Compute the test statistic.

(iv) The null hypothesis is to be tested at 95% confidence. Determine the critical value for this test.

5b. **The manager of a company believes that differences in sales performance depend upon the salesperson’s age. Independent samples of salespeople were taken and their weekly sales record is reported below.**

Below 30 years No. of Sales	Between 30 and 45 years No. of Sales	Over 45 years No. of Sales
24	23	30
16	17	20
21	22	23
15	25	25
19	18	34
26	29	36
	27	28

(i) State the null and alternative hypotheses.
(ii) At 95% confidence, test the hypotheses using Kruskal Wallis Test. (AU JAN 2018)

UNIT V: CORRELATION, REGRESSION AND TIME SERIES ANALYSIS

PART – A

1. **Define time series.**
A time series is an arrangement of statistical data in accordance with the time of occurrence in chronological order.

2. **Write angle between the regression lines**

$$\tan\theta = \left(\frac{1-r^2}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x{}^2+\sigma_y{}^2}$$

3. **When do you say two regression lines coincide with each other?**
When r = ±1 the two regression lines coincide .

4. **Differentiate between correlation and regression (APR/MAY 2018)**

Correlation analysis	Regression analysis
1. Correlation coefficient r between X and Y is a measure of linear relationship between X and Y	1. The regression coefficients are mathematical measures expressing the average relationship between the two variables
2. The correlation coefficient does not reflect upon the nature of variable	2. Regression coefficient reflect on the nature of variable
3. It is a relative measure and is independent of the units of measurement	3. Regression coefficients are absolute measures of finding out the relationship between two or more variables

5. **Write any two properties of regression co-efficient**
The coefficient of correlation is the geometric mean of the coefficients of regression
If one of the regression coefficients is greater than unity, then other is less than unity.

6. **Find the mean of x and y, given two regression lines are x+6y = 4 and 2x+3y = -1**
x + 6y = 4,2x + 3y = -1
2x + 12y = 8,2x +3y = -1 solving we get, x = -2, y = 1.

7. **The equations of the regression lines are given by 3x + y =10,**
3x + 4y = 12. Find the correlation coefficient between x and y
x on y : x = 10/3 – 1/3 y
b_{xy} = -1/3
y on x : y = 12/4 –3/4 x
b_{yx} = -3/4 , $r^2 = b_{xy}b_{yx}$ = (–1/3) (–3/4) = 1/4 = 0.25

8. **What are the basic components of Time series analysis? (AU JAN 2015)**
1.Secular Trend 2.Seasonal Variations 3.Cyclical Variations 4.Irregular Variations

9. **Define regression.**
Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

10. **Write the normal equations for the method of fitting of a parbolic curve.**
$y=a+bx+cx^2$ is the trend equation and the normal equations are
$\sum y=na+b\sum x+c\sum x^2$, $\sum xy=a\sum x +b\sum x^2+c\sum x^3$, $\sum x^2y=a\sum x^2 +b\sum x^3+c\sum x^4$

11. **What are the uses of regression analysis (AU MAY/JUNE 2019)**
   1. It is useful in economic analysis as regression equation can determine an increase in the cost of living index for a particular increase in general price level.
   2. It enables us to study the nature of relationship between the variables.

12. **Write the merits of the least squares method.**

1. This method gives the trend values for the entire time period
2. This method is a completely objective in character
3. This method can be used to forecast future trend because trend line establishes a functional relationship between the value and the time.

13. **If the tangent of the angle between the lines of regression y on x and x on y is 0.6 and $\sigma_x = \frac{1}{2}\sigma_y$.**

Find the correlation coefficient.

$$\tan\theta = \frac{\sigma_x\sigma_y}{\sigma^2{}_x + \sigma^2{}_y}\left[\frac{1-r^2}{r}\right],\ 0.6 = \frac{\frac{1}{2}\sigma_y\sigma_y}{\frac{1}{4}\sigma^2{}_y + \sigma^2{}_y}\left[\frac{1-r^2}{r}\right]$$

$$= \frac{\frac{1}{2}}{\frac{5}{4}}\left[\frac{1-r^2}{r}\right] => 1.5r = 1 - r^2, r = -2,0.5$$

14. **State any two properties of correlation coefficient.**

(i) The coefficient of correlation lies between -1 and +1.

(ii) The coefficient of correlation is independent of change of scale and origin of the variables X & Y.

15. **What are the various methods of studying trend?**

1. Graphic method 2. Method of semi-averages 3. Method of Moving averages
4. Method of Least squares.

16. **Write down the formula to calculate rank correlation coefficient .**

$\rho_s = 1 - \frac{6(\sum d_i^2)}{n(n^2-1)}$, $d_i = x_i - y_i$

17. **Briefly explain how a scatter diagram benefits the researcher? (AU MAY/JUNE 2014)**

The simplest device for studying correlation between two variables is a special type of dot chart called scatter diagram. In this method, the given data is plotted on a graph in the form of dots. The more the plotted points scatter over a chart, the lesser is the degree of relationship between the two variables. The nearer the points come to the line, the higher the degree of relationship. If the plotted points lie in a haphazard manner it shows the absence of any relationship between the variables.

18. **When do we say the variables are positively correlated, negatively correlated and uncorrelated.**

(i) If r=1 then there is a perfect positive correlation.

(ii) If r= -1 then there is a perfect negative correlation.

(iii) If r=0 then the variables are uncorrelated.

19. **Mention the two mathematical models for a time series.**

1. Additive model – This model assumes that the four components of the time series Trend, seasonal, cyclical and irregular variations are independent of each other.
2. Multiplicative model - This model assumes that the four components of the time series are interdependent.

20. **State the limitations of Method of Moving averages.**

(1) Trend values cannot be calculated for all the years that is, some years will be left out in the beginning and in the end.

(2) The period of moving average has to be chosen with great care.

(3) This method cannot be used for forecasting.

PART B

1a. **Calculate the coefficient of correlation between X and Y , using the following data : (AU APR 2018)**

X	**1**	**3**	**5**	**7**	**8**	**10**
Y	**8**	**12**	**15**	**17**	**18**	**20**

1b. **Fit a second degree polynomial equation for the following data**

X	1976	1977	1978	1979	1980	1981	1982	1983	1984
Y	50	65	70	85	82	75	65	90	95

2. **Given below are the figures of production (in thousand quintals) of a sugar factory.**

Year	1974	1975	1976	1977	1978	1979	1980
Production	77	88	94	85	91	98	90

Fit a straight line by the least squares method and tabulate the trend values.

3a. **Find the two regression lines using the data below: (AU NOV/DEC 2018)**

X	7	4	8	6	5
Y	6	5	9	8	2

3b. **The following data on production (in ‘000 units) of a commodity from the year 2006-2012. Fit a straight line trend and forecast for the year 2020 (AU NOV/DEC 2017)**

Year	2006	2007	2008	2009	2010	2011	2012
Production	6	7	5	4	6	7	5

4a. **Explain the basic components of Time series analysis.**

4b. **The monthly water consumption in thousand gallons in a hostel for five years is given below. Calculate the seasonal indices by the method of simple averages**

Year	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
1979	25	23	21	18	15	20	21	25	22	24	32	35
1980	27	25	23	20	17	22	23	27	24	26	35	33
1981	32	31	30	27	25	27	29	30	30	32	41	38
1982	42	40	38	36	34	37	38	40	38	43	52	48
1983	57	50	52	46	49	46	49	55	50	59	64	63

5a. **The following table gives the profits of a concern for 5 years ending 1983. Fit an exponential curve for the following data (AU MAY/JUNE 2019)**

Year	1979	1980	1981	1982	1983
Profits	1.6	4.5	13.8	40.2	125.0

5b. **The quarterly sales (in thousands of copies) for a specific education software over the past three years are given in the following table.**

	2003	2004	2005
Quarter 1	170	180	190
Quarter 2	111	96	120
Quarter 3	270	280	290
Quarter 4	250	220	223

(i) Compute the four seasonal factors (Seasonal Indexes). Show all of your computations.

(ii) The trend for these data is Trend = 174+4t (t represents time, where t=1 for Quarter 1 of 2003 and t=12 for Quarter 4 of 2005). Forecast sales for the first quarter of 2006 using the trend and seasonal indexes. Show all of your computations.

25/07/23

Reg. No. : □□□□□□□□□□□□

Question Paper Code : 10191

M.B.A. DEGREE EXAMINATIONS, APRIL/MAY 2023.

First Semester

BA 4101 – STATISTICS FOR MANAGEMENT

(Regulations 2021)

Time : Three hours

Maximum : 100 marks

Approved Table may be permitted.

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. Define conditional probability.

2. What is a Binomial Experiment?

3. What is estimation?

4. State central limit theorem.

5. What is a null hypothesis?

6. What is Type I error?

7. What are the advantages of non - parametric tests?

8. When do you use a chi square test?

9. What is a regression model?

10. What is a random error in a regression equation?

PART B — (5 × 13 = 65 marks)

11. (a) In a bolt factory machines A, B and C manufacture respectively 25%, 35% and 40% of the total. Of their output, 5%, 4% and 2% are defective bolts. A bolt is drawn at random and found to be defective, What is the probability that it was manufactured by machine C?

Or

(b) Hupper Corporation produces many types of soft drinks, including Orange Cola. The filling machines are adjusted to pour 12 ounces of soda into each 12-ounce can of Orange Cola. However, the actual amount of soda poured into each can is not exactly 12 ounces; it varies from can to can. It has been observed that the net amount of soda in such a can has a normal distribution with a mean of 12 ounces and a standard deviation of 0.015 ounce. (i) What is the probability that a randomly selected can of Orange Cola contains 11.97 to 11.99 ounces of soda? (ii) What percentage of the Orange Cola cans contain 12.02 to 12.07 ounces of soda?

12. (a) A publishing company has just published a new college textbook. Before the company decides the price at which to sell this textbook, it wants to know the average price of all such textbooks in the market. The research department at the company took a sample of 36 comparable textbooks and collected information on their prices. This information produced a mean price of Rs.145 for this sample. It is known that the standard deviation of the prices of all such textbooks is Rs.35 and the population of such prices is normal. (i) What is the point estimate of the mean price of all such college textbooks? (ii) What is the margin of error at 90% confidence interval. (iii) Construct a 90% confidence interval for the mean price of all such college textbooks.

Or

(b) According to a survey conducted by Pew Research Center in June 2009, 44% of people aged 18 to 29 years said that religion is very important to them. Suppose this result is based on a sample of 1000 people aged 18 to 29 years. (i) What is the point estimate of the corresponding population proportion? (ii) What is the margin of error at 99% confidence level. (iii) Find, with a 99% confidence level, the percentage of all people aged 18 to 29 years who will say that religion is very important to them.

13. (a) A test was given to students of two groups A and B to test whether there is any difference in the learning ability. Sixteen students of Group A took the test and their average score was found to be 55.8 with the standard deviation of 5.7. Nine students of Group B took the test and their average score was found to be 59.3 with the standard deviation of 4.3. At 5% significance level, can you conclude that the learning ability of the students of Group A and B are different? Assume that the population standard deviations are equal.

Or

(b) To test the significance of variation in the retail prices of a commodity in 3 cities, Mumbai, Kolkata and Delhi, 4 shops were chosen at random in each city and the prices are given as follows.

Mumbai :	16	8	12	14
Kolkata :	14	10	10	6
Delhi :	4	10	8	8

Are the prices in these cities different. Assume that the population is normally distributed with equal variance and the data collected randomly.

14. (a) A college administration is interested in checking whether the application for admission arrive randomly with respect to the gender. The gender of 25 consecutively arriving application were found to arrive in the following order, where M denotes a male applicant and F denotes a female applicant.

M,F,M,M,F,F,F,M,F,M,M,M,F,F,F,F,M,M,M,F,F,M,F,M,M

Can you conclude that the applications for admission arrive randomly with respect to gender at 95% confidence level?

Or

(b) A dairy agent wants to test a hormone that may increase cow's milk production. Some members of the group fear that the hormone could actually decrease production. So a matched pairs test is arranged. 30 cows were given the hormone and their milk production is recorded for 4 weeks. Each of these 30 cows is matched with another cow of similar size, age and prior record of milk production. This second group of 30 cows do not receive the hormone. The milk production of these cows were recorded for the same period of time. In the 19 of these 30 pairs, the cow taking the hormone produced more milk, in 9 of the pairs, the cow taking the hormone produced less and in 2 of the pairs, there was no difference. Using 5% level of significance, can you conclude that the hormone changes the median milk production of such cows?

15. (a) The following table gives indices of industrial production and the number of unemployed people in a state in lakhs. Check whether industrial production and the number of unemployed people are related by computing the correlation coefficient.

Index of Production :	100	102	105	107	105	112	103	99
Number Unemployed :	15	12	13	11	12	12	19	26

Or

(b) Find the line of best fit for the following data. AD indicates cost of advertisement and SR indicates Sales Revenue. Calculate the error of estimate.

AD	21	22	23	24	22	20	20	21	22	25
SR	115	125	120	133	142	150	155	135	125	120

PART C — (1 × 15 = 15 marks)

16. (a) Connie Rodrigues, the Dean at Midstate College is wondering about the grade distributions at the school. She has heard that the GPAs in Business School are about 0.25 lower than those in college of Arts and Science. A quick random sampling produced the following GPAs.

Business	2.86	2.77	3.18	2.8	3.14	2.87	3.19	3.24
Arts and Science	3.35	3.32	3.36	3.63	3.41	3.37	3.45	3.43

Business	2.91	3	2.83		
Arts and Science	3.44	3.17	3.26	3.18	3.41

Assuming that the corresponding populations are normally distributed, do these data indicate that there is a factual basis for the grumblings? State and test the appropriate hypothesis at 98% confidence level.

Or

(b) A brand manager is concerned that her brand's share may be unevenly distributed throughout the country. In a survey in which the country was divided into 4 geographic regions, a random sampling of 100 consumers in each region was surveyed with the following results. In the North East region, 40 purchased the brand and the rest did not purchase. In the North West region, 55 purchased the brand and the rest did not purchase. In the South East region, 45 purchased the brand and the rest did not purchase. In the South West region, 50 purchased the brand and the rest did not purchase. At $\alpha = 0.05$, Use Chi Square test to check whether the brand share is the same across the four regions.

Reg. No. : □□□□□□□□□□□□

Question Paper Code : 30076

M.B.A. DEGREE EXAMINATIONS, NOVEMBER/DECEMBER 2022.

First Semester

BA 4101 – STATISTICS FOR MANAGEMENT

(Regulations 2021)

Time : Three hours

Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. State the Baye's theorem.

2. Find $p(x=2)$ if $\lambda=5.2$ for a Poisson distribution.

3. List the two properties of sampling distribution of the mean when the population is normally distributed.

4. What are the applications of central limit theorem?

5. Define Type I and Type II errors.

6. What are the five steps of a test of hypothesis using the critical value approach?

7. Define Kolmogorov-Smirnov Test. Write its advantages.

8. Explain the chi-square distribution. What is the parameter of such a distribution?

9. Distinguish between correlation and regression.

10. What do you mean by standard error of estimate? Write its equation.

PART B — (5 × 13 = 65 marks)

11. (a) A statistical experiment has 11 equally likely outcomes that are denoted by $a, b, c, d, e, f, g, h, i, j$ and k. Consider three events : $A = \{b, d, e, f\}$, $B = \{a, c, f, j\}$ and $C = \{c, g, k\}$.

(i) Are events A and B independent events? What about events A and C?

(ii) Are events A and B mutually exclusive events? What about B and C?

(iii) What are the complements of events A, B and C respectively? What are their probabilities?

Or

(b) At the Express House Delivery Service, Providing high quality service to customers is the top priority of the management. The company guarantees a refund of all charges if a package it is delivering does not arrive at its destination by the specified time. It is known that from past data, 2% of the packages mailed through this company do not arrive at their destinations within the specified time. Suppose a corporation mails 10 packages through Express House Delivery Service on a certain day.

(i) Find the probability that exactly one of these 10 packages will not arrive at its destination within the specified time.

(ii) Find the probability that atmost one of these 10 packages will not arrive at its destination within the specified time.

12. (a) (i) In a sample of 25 observations from a normal distribution with mean 98.6 and standard deviation 17.2

(1) What is $P(92 < \bar{x} < 102)$?

(2) Find the corresponding probability given a sample of 36. (7)

(ii) It is said that happy and healthy workers are efficient and productive. A company that manufactures exercising machines wanted to know the percentage of large companies that provide on-site health club facilities. A random sample of 240 such companies showed that 96 of them provide such facilities on site.

(1) What is the point estimate of the percentage of all such companies that provide such facilities on site?

(2) Construct a 97% confidence interval for the percentage of all such companies that provide such facilities on site. What is the margin error for this estimate? (6)

Or

(b) (i) In a large city, 88% of the cases of car burglar alarms that go off are false. Let $\overline{p}$ be the proportion of false alarms in a random sample of 80 cases of car burglar alarms that go off in this city. Calculate the mean and standard deviation of $\overline{p}$, also describe the shape of its sampling distribution. (6)

(ii) Suppose a total of 789,654 families live in a particular city and 563,282 of them own homes. A sample of 240 families is selected from this city, 158 of them own homes. Find the proportion of families who own homes in the population and in the sample. (7)

13. (a) Two independent samples of observations were collected. For the first sample of 60 elements, the mean was 86 and the standard deviation 6. The second sample of 75 elements had a mean of 82 and a standard deviation of 9.

Compute the estimated standard error of the difference between the two means.

Using $\alpha = 0.01$, test whether the two samples can reasonably be considered to have come from populations with the same mean.

Or

(b) Given a sample mean 83, a sample standard deviation of 12.5 and a sample size of 22, test the hypothesis that the value of the population mean is 70 against the alternative that it is more than 70. Use the 0.025 level of significance.

14. (a) Fit a Poisson distribution for the following distribution and also test the goodness of fit.

x:	0	1	2	3	4	5
f:	142	156	69	27	5	1

Or

(b) Use Kruskal-Wallis test to test the difference in mean among the three samples.

Sample 1 :	99	64	101	85	79	88	97	95	90	100
Sample 2 :	83	102	125	61	91	96	94	89	93	75
Sample 3 :	89	98	56	105	87	90	87	101	76	89

15. (a) The following are ratings of aggressiveness (X) and amount of sales in the last year (Y) for 8 sales people. Calculate the rank correlation between two measures. Use 0.10 significance level.

X:	30	17	35	28	42	25	19	29
Y:	35	31	43	46	50	32	33	42

Or

(b) Calculate the regression coefficient and obtain the lines of regression for the following data :

X:	1	2	3	4	5	6	7
Y:	9	8	10	12	11	13	14

Obtain an estimate of Y which correspond to the value $X = 5.4$.

PART C — (1 × 15 = 15 marks)

16. (a) A study compared the effects of four 1 month point of purchase promotions on sales. The unit sales for five stores using all four promotions in different months follow :

Free sample :	78	87	81	89	85
One-pack gift :	94	91	87	90	88
Cents off :	73	78	69	83	76
Refund by mail :	79	83	78	69	81

(i) Compute the mean unit sales for each promotion and then determine the grand mean.

(ii) Estimate the population variance using the between – column = variance.

(iii) Estimate the population variance using the within-column variance computed from the variance within the samples

(iv) Calculate the F-ratio. At the 0.01 level of significance, do the promotions produce different effects on sales?

Or

(b) Cost accountants often estimate overheads based on the level of production. At the Standard Knitting Co., they have collected information on overhead expenses and units produced at different plants, and want to estimate a regression equation to predict future overhead.

Overhead :	191	170	272	155	280	173	234	116	153	178
Units :	40	42	53	35	56	39	48	30	37	40

(i) Develop the regression equation for the cost accountants.

(ii) Predict overhead when 50 units are produced.

(iii) Calculate the standard error of estimate.