GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

Seshadri Rao Knowledge Village, Gudlavalleru – 521 356.

Department of Computer Science and Engineering

HANDOUT

on

PROBABILITY & STATISTICS

Vision

To be a Centre of Excellence in computer science and engineering education and training to meet the challenging needs of the industry and society

Mission

- To impart quality education through well-designed curriculum in tune with the growing software needs of the industry.
- To be a Centre of Excellence in computer science and engineering education and training to meet the challenging needs of the industry and society.
- To serve our students by inculcating in them problem solving, leadership, teamwork skills and the value of commitment to quality, ethical behavior & respect for others.
- To foster industry-academia relationship for mutual benefit and growth

Program Educational Objectives

PEO1 : Identify, analyze, formulate and solve Computer Science and Engineering problems both independently and in a team environment by using the appropriate modern tools.

PEO2 : Manage software projects with significant technical, legal, ethical, social, environmental and economic considerations.

PEO3 : Demonstrate commitment and progress in lifelong learning, professional development, leadership and Communicate effectively with professional clients and the public

HANDOUT ON PROBABILITY & STATISTICS

Class & Sem. : II B.Tech – II Semester Year : 2019-20

Branch : IT Credits : 3

==========

1. **Brief History and Scope of the Subject**

   The History of Foundations of Mathematics involve non classical logics and constructive mathematics. Mathematical Foundations of Computer Science is the study of mathematical structures that are fundamentally discrete rather than continuous. Research in Discrete Structures increased in the latter half of 20^{th} centenary partly due to development of digital computers, Which operate in Discrete steps and store data in discrete bits. Graph Theory is study of, Mathematical Structures used to model pair wise relations between objects from a certain collection. This course is useful in study and describing objects and problems in computer science such as computer algorithm, programming languages, Cryptography, Automated theorem proving and software development.

2. **Pre-Requisites**
   - Mathematics background such as set theory, basics in probability & basics in statistics.

3. **Course Objectives:**
   - To impart the concepts of probability and statistics.
   - To disseminate the knowledge on sampling theory and principles of hypothesis testing.
   - To introduce the correlation coefficient and lines of regression.

4. **Course Outcomes:**

   Upon successful completion of the course, the students will be able to
   - use the concepts of probability in different real time problems.
   - apply probability distribution in appropriate scenario.
   - find confidence intervals for estimating population parameters.
   - apply a range of statistical tests appropriately.
   - measure correlation between variables and obtain lines of regression.

5. **Program Outcomes:**

   Engineering Graduates will be able to:

   1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
   2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
   3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with

appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

6. Mapping of Course Outcomes with Program Outcomes:

	1	2	3	4	5	6	7	8	9	10	11	12
CO1	2	2										
CO2		3										
CO3	3	3										
CO4	3	3										
CO5	2	2										

MAPPING OF COURSE WITH PO'S

SUBJECT NAME	1	2	3	4	5	6	7	8	9	10	11	12
PROBABILITY & STATISTICS	3	3	-	-	-	-	-	-	-	-	-	-

7. Prescribed Text Books

a. Dr. T. K. V. Iyengar, Dr. B. Krishna Gandhi, S. Ranganatham and Dr. M.V. S. S. N. Prasad, Probability and Statistics, S. Chand & Company Ltd., New Delhi.

b. Miller, John E. Freund, Probability and Statistics for Engineers, PHI, Delhi.

8. Reference Text Books

a. S.C. Gupta & V.K. Kapoor, Fundamentals of Mathematical Statistics, S.Chand & Company Ltd., New Delhi.

b. B.V. Ramana, Engineering Mathematics, 4th Edition, Maitrey Printers Pvt. Ltd., 2009, India.

9. URLs and Other E-Learning Resources

So net CDs & IIT CDs on some of the topics are available in the digital library.

10. Digital Learning Materials:

a. www.mathworld .wolfram.com
b. www.social research methods. net/kb/samprob. php
c. www.fourmilabch/rpkp/experiments/statistics.html
d. www.Hypothesis –Testing.html
e. http://quizlet.com
f. www.probabilitycourse. com

10. Digital Learning Meterial:

- http://www.socr.ucla.edu
- www.statlect.com
- www.stat.ucla.edu

11. Lecture Schedule:

Topic	No. of Periods	
	Theory	Tutorial
UNIT –1: Probability		
Introduction to probability	1	1
Simple problems	1	
Addition theorem-problems	1	
Conditional and multiplication theorem-problem	1	
Independent Events- Problems	1	1
Baye's theorem-problems	1	
Applications.	1	
Random variables: Discrete Random variable, Pmf ,distribution function	1	
Problems on DRV-Mean, Variance, different probabilities	1	1
Problems on DRV	1	
Continuous random variable, pdf, Distribution function	1	
Problems on CRV- Mean, Variance, different probabilities	1	
Problems on CRV	1	
UNIT – 2: Standard Probability Distributions		1
Binomial distribution: introduction - mean and variance	1	
Problems on Binomial distribution	2	
Poisson distribution : mean and variance	2	
Normal distribution – Properties	1	1
Area property Problems	2	

Applications of uniform distribution	1	
Applications of exponential distribution	1	
UNIT – 3: Sampling Distributions		
Population, samples, parameter, statistic, random sample, sampling distribution, standard error.	1	1
Sampling distribution of mean -problems on with replacement	2	
Sampling distribution of mean- Problems on without replacement	1	
Sampling distribution of difference and sums – problems	1	
Sampling distribution of difference and sums – problems	1	
Introduction to estimation – point estimation – results- Problems	1	1
Interval estimation: confidence Intervals for means –problems	1	
Confidence interval for proportions -problems	1	
UNIT – 4: Testing Of Hypothesis (Large Samples)	1	
Test of hypothesis- simple, composite hypotheses, Null hypothesis and alternative Hypothesis, Test statistic. Type I & Type 2 errors in sampling.	1	
L.O.S – one tail and two tail tests, degrees of freedom, procedure of testing of hypothesis.	1	1
Test of significance of single mean –large samples- problems.	2	
Test of significance of two mean –large samples- problems.	1	
Problems.	1	
Hypothesis concerning one proportion-problems.	1	1
Problems.	1	
Hypothesis concerning two proportions-problems.	1	
Problems.	1	1
UNIT – 5: Testing Of Hypothesis (Small Samples)		
Tests of significance: students t-test – means	1	
Problems on t-test	1	1
Tests of significance: students t-test – two means	1	
Paired t-test -problems	1	
F-test-problems	1	
Analysis of r x c tables – chi- square test for independence	1	1
Chi- square test for Goodness of fit	1	
Chi- square test for Goodness of fit using Poisson distribution	1	
UNIT – 6 Correlation-Regression And Queueing Theory		
Simple correlation ,types of correlation, correlation co-efficient	2	
Problems on correlation coefficient	1	1
rank correlation -problems	1	
Linear regression and its properties	1	
TOTAL	**56**	**14**

12. Seminar Topics

- Probability
- Probability Distributions
- Sampling Distributions
- Significance Tests
- Correlation and Regression

LEARNING MATERIAL

UNIT 1 – Probability

Objectives:

(a) To understand the concepts of probability and statistics.
(b) To know sampling theory and principles of hypothesis testing
(c) To appreciate Queuing theory and models.

Syllabus :

(a) Axioms of probability (Non-negativity, Totality, and Additivity)
(b) Conditional and Unconditional probabilities (Definitions and simple problems)
(c) Additive law of probability (simple applications)
(d) Multiplicative law of probability (simple applications)
(e) Baye's Theorem (without proof and applications)
(f) Concept of a Random variable (one dimensional case definition only and simple examples)
(g) Types of random variables (Discrete and Continuous cases)
(h) Probability mass function and probability density function – their properties (without proofs)
(i) Distribution Function and its properties (without proofs)
(j) Evaluation of mean and variance (problems)

Learning Outcomes:

Students will be able to

(a) understand the usage of axioms of probability
(b) apply various laws of probability (like additive, multiplicative, and Baye's) in real-life problems
(c) distinguish between discrete random variable (DRV) and continuous random variable (CRV)
(d) understand the complexity in finding the mean and the variance of DRV and CRV.

Learning Material

Terminology associated with Probability Theory:

Random experiment: If an experiment or trial can be repeated any number of times under similar conditions and it is possible to enumerate the total number of outcomes, but an individual outcome is not predictable, such an experiment is called a random experiment. For instance, if a fair coin is tossed three times, it is possible to enumerate all the possible eight sequences of head (H) and tail (T). But it is not possible to predict which sequence will occur at any occasion.

Outcome**:** A result of an experiment is termed as an outcome. Head (H) and tail (T) are the outcomes when a fair coin is flipped.

Sample Space: Each conceivable outcome of a random experiment under consideration is said to be a sample point. The totality of all conceivable sample points is called a sample space. In other words, the list of all possible outcomes of an experiment is called a sample space. For example, the set $\{HH, HT, TH, TT\}$ constitutes a sample space when two fair coins tossed at a time.

(i) *Discrete Sample Space:* A sample space which consists of countably finite or infinite elements or sample points is called discrete sample space. It is abbreviated as DSS.

(ii) *Continuous Sample Space:* A sample space which consists of continuum of values is called continuous sample space. It is abbreviated as CSS.

***Event*:** Any subset of the sample space is an event. In other words, the set of sample points which satisfy certain requirement(s) is called an event. For example, in the event, there are exactly two heads in three tossing's of a coin, it would consist of three points (H, H, T), (H, T, H), and (T, H, H). Each point is called an event. i.e. an outcome which further cannot be divided is called an event. Events are classified as:

***Elementary Event*:** An event or a set consists only one element or sample point is called an elementary event. It is also termed as simple event.

***Complementary Event*:** Let A be the event of S. The non-occurrence of A and contains those points of the sample space which do not belong to A.

***Exhaustive Events*:** All possible events in any trial are known as exhaustive events. In tossing a coin, there are two exhaustive elementary events namely, head and tail.

Equally Likely Events: Events are said to be equally when there is no reason to expect anyone of them rather than anyone of the others in a single trial of the random experiment. In other words, all the sample units or outcomes of sample space are having equal preference to each other, then the events are said to be equally likely events. In a tossing a coin, the outcomes head (H) and tail (T) are equally likely events.

Mutually Exclusive Events: Events A and B are said to be mutually exclusive events if the occurrence of A precludes the occurrence of B and vice-versa. In other words, if there is no sample point in A which is common to the sample point in B, i.e. $A \cap B = \phi$, the events A and B are said to be mutually exclusive. For example, if we flip a fair coin, we find either H or T in a trial, but not both. i.e. happening of H that prevents the happening of T in a trial, then H and T are mutually exclusive events. (No two events can happen simultaneously in a trial, such events are mutually exclusive.)

Independent events: Two events A and B are said to be independent if the occurrence of A has no bearing on the occurrence of B i.e. the knowledge that the event A has occurred gives no information about the occurrence of the event B.

Formally, two events A and B are independent if and only if,

$$P(A \cap B) = P(A)\,P(B).$$

For example, a bag contains balls of two different colours say, red and white. The two balls are drawn successively. First a ball is drawn from one bag and replaced after noting its color. Let us

presume that it is white and is denoted by the event A. Another ball is drawn from the same bag and its colour is noted. Let this event noted by the event B. The result of the second drawn is not influenced by the first drawn. Hence the events A and B are said to be independent.

Various definitions of probability are 1. Classical definition of probability (or Mathematical definition of probability) 2. Statistical definition of probability (or Empirical definition of probability) 3. Axiomatic approach to probability.

The classical definition of probability breaks down when we do not have a complete priori analysis i.e. when the outcomes of the trial are not equally or when the total number of trials is infinite or when the enumeration of all equally likely events is not possible. So the necessity of the statistical definition of probability arises.

The statistical definition of probability, although is of great use from practical point of view, is not conductive for mathematical approach since an actual limiting number may not really exist. Hence another definition is thought of based on axiomatic approach. This definition leads to the development of calculus of probability.

- P(E) = Favourable number of cases / Total cases

- Limits of probability $0 \leq \text{P (E)} \leq 1$

Axiomatic approach to Probability: A real valued function $p(x):S(x) \to (0,1)$ is called a probability function which satisfies the following rules or statements technically termed as axioms, where S is a sample space, x is a result of an experiment ranges from $-\infty$ to ∞.
Axiom 1: (Non-negativity) For any event E of S, $P(E) \geq 0$.
Axiom 2: (Totality) S be a sample space of an experiment and P(S) = 1.
Axiom 3: (Additive) Suppose E_1 and E_2 be mutually exclusive events of S, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

For example, police department needs new tires for its patrol cars and the probabilities are 0.17, 0.22, 0.03, 0.29, 0.21, and 0.08 that it will buy Uniroyal tires, Goodyear tires, Michelin tires, General tires, Goodrich tires or Armstrong tires. Then the probabilities that the combinations of (Goodyear, Goodrich), (Uniroyal, General, Goodrich), (Michelin, Armstrong), and (Goodyear, General, Armstrong) tires respectively are 0.43, 0.67, 0.11, and 0.59 respectively.

Note: Axioms of probability do not determine probabilities. But the axioms restrict the assignments of probabilities in a manner that enables us to interpret probabilities as relative frequencies without inconsistencies.

Unconditional Probability:
The individual probabilities of the events of A and B are termed as unconditional probabilities. i.e. unconditional probabilities are the probabilities which are not influenced by other events in the sample space, S. These are also termed as priori probabilities.

Result : If A is any event in s, then $P(\overline{A}) = 1 - P(A)$

Result : For any two events A and B then

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

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

Additive Law of Probability:

Statement: If A and B are any two events of a sample space S and are not disjoint then

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

Proof:

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

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

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

$$= P(A) + [P(\overline{A} \cap B) + P(A \cap B) - P(A \cap B)]$$

$$= P(A) + P[(\overline{A} \cap B) \cup (A \cap B)] - P(A \cap B)$$

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

Result : If A and B are disjoint events then $P(A \cup B) = P(A) + P(B)$

Example: A card is drawn from a well shuffled pack of cards. What is the probability that it is either a spade or an ace?

Solution: We know that (WKT), a pack of playing cards consists 52 in number. i.e. the sample space of pack of cards, n(S) = 52.

Let A denoted the event of getting a spade and B denotes the event of getting an ace. Then the probability of the event of getting either a spade or an ace is

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

Since all the cards are equally likely, mutually exclusive, we have

$$P(A) = \frac{13}{52},\ P(B) = \frac{4}{52},\ P(A \cap B) = \frac{1}{52}$$

By addition theorem of probability,

$$P(A \cup B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{4}{13}.$$

Conditional Probability: In many situations arises in our day to day life about the occurrence of an event A (for instance, getting treatment) is influenced by the occurrence of the event B (availability of doctor) and the event is known a conditional event, denoted by $A \mid B$ and hence the probability of the conditional event is known as 'conditional probability' and is denoted by $P(A \mid B)$.

Definition: Let A and B be two events. The conditional probability of event B, if an event A has occurred, is defined by the relation,

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)},\ \textit{if}\ P(A) > 0.$$

i.e. the conditional probability of the event B is the ratio of the probability of the joint occurrence of the events A and B to the unconditional probability of the event A.

Similarly, we can define $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$, *if* $P(B) > 0$.

Example: In a group consisting of men and women are equal in number. 10% of the men and 45% of the women are unemployed. If a person is selected randomly from the group then find the probability that the person is an unemployed.

Solution: Since the men and women are equal in number in a group, we take

$$P(M) = 1/2 \text{ and } P(W) = 1/2$$

Let E be the event of employed person. Then $\bar{E}$ be the event of unemployed.

Then we have, $P(E \mid M) = 10\% = 0.10$, $P(E \mid W) = 40\% = 0.45$

implies, $P(\bar{E} \mid M) = 0.90$, $P(\bar{E} \mid W) = 0.55$

The probability that the person (either male or female) is an unemployed is

$$P(\bar{E}) = P(M)\,P(\bar{E} \mid M) + P(W)\,P(\bar{E} \mid W)$$

$$= \frac{1}{2}(0.90) + \frac{1}{2}(0.55) = 0.725$$

Example: Two marbles are drawn in succession from a box containing 10 red, 30 white, 20 blue and 15 orange marbles with replacement being made after each draw. Find the probability that (i) both are white (ii) first is red and second is white.

Solution: From the given information, we noticed that

the number of marbles in the box = 75.

i. Let us define E_1 be the event of 1st drawn marble is white and E_2 be the event of 2nd drawn marble is also white.

Since we are using with replacement to select marbles in succession, we have

$$P(E_1) = \frac{30}{75} \text{ and } P(E_2) = \frac{30}{75}$$

Therefore, the probability that both marbles are white is

$$P(E_1 \cap E_2) = P(E_1)\,P(E_2 \mid E_1) = \frac{30}{75}.\frac{30}{75} = \frac{4}{25}$$

ii. Let us define E_1 be the event of 1st drawn marble is red and E_2 be the event of 2nd drawn marble is white.

Since we are using with replacement to select marbles in succession, we have

$$P(E_1) = \frac{10}{75} = \frac{2}{15} \text{ and } P(E_2) = \frac{30}{75} = \frac{2}{5}$$

Therefore, the probability that the first draw marble is red and the second draw marble is white is

$$P(E_1 \cap E_2) = P(E_1)\,P(E_2 \mid E_1) = \frac{2}{15}.\frac{2}{5} = \frac{4}{75}.$$

Multiplicative Law of Probability:

Statement: For any events A and B in the sample space S, we have

$$P(A \cap B) = P(A)\,P(B \mid A),\ P(A) > 0$$
$$= P(B)\,P(A \mid B),\ P(B) > 0$$

Where $P(B \mid A)$ is the conditional probability of B provided A has already happened and $P(A \mid B)$ is the conditional probability of A provided B has already happened.

Result: If A and B are independent events then P(A ∩ B)= P (A) P(B)

Result: If A1,A2……..An are n independent events then probability of happening of at least one of the event= 1- probability of none of the events happening

Baye's Theorem:

Statement: Suppose $E_1, E_2, ..., E_n$ be 'n' mutually exclusive events in S with $P(E_i) \neq 0;\ i = 1, 2, ..., n.$ Let A be any arbitrary event which is a subset of S and $P(A) > 0.$

Then, we have $P(E_i \mid A) = \dfrac{P(E_i)\,P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\,P(A \mid E_i)}$, $i = 1, 2, ..., n$.

Where $P(E_i)'s$ are called 'a priori probabilities', $P(A \mid E_i)'s$ are called 'likelihoods' and $P(E_i \mid A)'s$ are called 'posterior probabilities'.

Note: $\sum_{i=1}^{n} P(E_i)\,P(A \mid E_i)$ = P(A) is called Total probability

Example: Four computer companies A, B, C and supply transistors to a company. From previous experience, it is known that the probability of the transistors being bad if it comes from A is 40%, from B is 2%, from C is 5% and from D is 1%. The probabilities of picking supplier A is 20%, B is 30%, C is 10% and D is 40%.

(i) Find the probability that a transistor chosen at random is bad.

(ii) Find the probability that the transistor comes from company A, given that the transistor is bad.

Sol: Probabilities of picking suppliers A,B,C and D are $P(E_1) = 0.2$, $P(E_2) = 0.3$, $P(E_3) = 0.1$, and $P(E_4) = 0.4$ respectively

Getting *suppliers, A,B,C and D when picked are events E1,E2,E3 and E4 respectively*

D is the event bad

Given $P(D/E_1) = 0.4$, $PD/(E_2) = 0.02$, $P(D/E_3) = 0.05$, $P(D/E_4) = 0.01$

(i) $P(D) = P(E_1)\,P(D/E_1) + P(E_2)\,P(D/E_2) + P(E_3)\,P(D/E_3) + P(E_4)\,P(D/E_4)$
$= 0.895$

(ii) $P(E_1/A) = P(E_1)\,P(D/E_1) / P(D) = 0.893$

Concept of Random Variable:

A random variable X is a real function of the events of a given sample space S. Thus for a given experiment defined by a sample space S with events s, the random variable is a function of s. It is denoted by X(s). A random variable X can be considered to be a function that maps all events of the sample space into points on the real axis.
For example, an experiment consists of tossing two coins. Let the random variable be a function X chosen as the number of heads shown. So X maps the real numbers of the event "showing no head" as zero, the event "any one is head" as one and "both heads" as two. Therefore, the random variable is X = $\{0,1,2\}$. The elements of the random variable X are $x_1 = 0,\ x_2 = 1,\ and\ x_3 = 2$.

Types of Random Variable:

Random variables are classified into:

1. *Discrete Random Variable* (DRV): A random variable X which is defined on the discrete sample space is called discrete random variable.
   For example, consider a discrete sample space $S = \{1,2,3,4\}$. Let us define X = S^2 be a random variable. Then discrete values of S map to discrete values of X as $\{1,4,9,16\}$. The probabilities of the random variable x are equal to the probabilities of set S because of the one-to-one mapping of the discrete points.
   Let X be a discrete random variable with integer events $X = \{x_1, x_2, ..., x_n\}$. The probability of X at any event is a function of x_i and is given by
   $$P(X = x_i) = p(x_i)\ ,\ i = 1,2,3,.....$$
   This function is called probability mass function and is abbreviated as p.m.f. (pmf).

Properties of probability mass function:
Consider a discrete random variable X in a sample space with infinite number of possible outcomes, that is, $X = \{x_1, x_2, ...\}$. If the probability of X, $p(x_i)$, $i = 1,2,3,....$ satisfies the following properties then the function $p(x)$ is called probability mass function.

(i) $p(x_i) \geq 0,\ \forall i$ (ii) $\sum_{i=1}^{\infty} p(x_i) = 1$

2. *Continuous Random Variable* (CRV): A random variable X which is defined on the continuous sample space is called continuous random variable.
   Temperature, time, height and weight over a period of time etc. are examples of CRV.
   The probability density function of a random variable X is defined as the variable X falls in the infinitesimal interval $\left[x - \frac{dx}{2}, x + \frac{dx}{2}\right]$ such that $P\left(x - \frac{dx}{2} \leq X \leq x + \frac{dx}{2}\right) = f(x)dx$,
   $$i.e.\ f(x) = \underset{\Delta t \to 0}{Lt} \frac{P\left(x - \frac{dx}{2} \leq X \leq x + \frac{dx}{2}\right)}{dx}$$

Where $f(x)$ is called the probability density function of a random variable X and the continuous curve y = f(x) is called probability density curve.

Properties of probability density function:
The continuous curve y =f(x) satisfies the following properties, then the function f(x) is called probability density function of a random variable and is abbreviated as p.d.f. (pdf).

(i) $f(x) \geq 0, \forall x,$ (ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

3. *Mixed Random Variable*: A random variable which is defined on both DSS and CSS partially, then the random variable is said to be a mixed random variable.

Probability distribution function:
Let X be a random variable. Then the probability distribution function associated with X is defined as the probability that the outcomes of an experiment will be one of the outcomes for which $X(s) \leq x, x \in R.$ That is, the function $F(x)$ is defined by
$F(x) = P(X \leq x) = P\{s : X(s) \leq x\}, -\infty < x < \infty$ is called the distribution function of X.
Sometimes it is also known as Cumulative Distribution function and is abbreviated as CDF.
Properties of cdf:
1 . If F is the distribution function of a random variable S and $a < b,$ then
(i) $P(a < X \leq b) = F(b) - F(a)$
(ii) $P(a \leq X \leq b) = P(X = a) + [F(b) - F(a)]$
(iii) $P(a < X < b) = [F(b) - F(a)] - P(X = b)$
(iv) $P(a \leq X \leq b) = [F(b) - F(a)] = P(X = b) + P(X = a)$
2. All distribution functions are monotonically increasing and lie between 0 and 1. That is, if F is the distribution function of the random variable X, then
(i) $0 \leq F(x) \leq 1$ i.e. F is bounded.
(ii) $F(x) < F(y)$ when $x < y.$
(ii) $F(-\infty) = 0$ and $F(\infty) = 0.$

Evaluation of mean and variance:

1. A random variable 'X' has the following probability functions:

x	0	1	2	3	4	5	6	7
P(x)	0	K	2K	2K	3K	K^2	$2K^2$	$7K^2+K$

(i) Determine 'K' (ii) Evaluate $P(X < 6), (PX \geq 6) \& Plo < x < 5$
(iii) Mean (iv) Variance

Sol: $Since \sum_{x=0}^{7} P(x) = 1$
K = 1/10 = 0.1
P(X<6)=P(X=0)+P(x=1)+……..+P(x=5)
=0.81

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

$Mean\,\mu = \sum_{i=0}^{7} P_i x_i$

= 3.66 (K=1/10)

Variance $= \sum_{i=0}^{7} P_i x_i^{\,2} - \mu^2$

=3.4044

2. The probability density f(x) of a continuous random variable is given by $f(x) = C.e^{-|x|}$, $-\infty < x < \infty$. S.T C = ½ and find (i) mean (ii) variance of the distribution. Also find $P(0 \le X \le 4)$.

Sol: We have $\int_{-\infty}^{\infty} f(x)dx = 1$

$\Rightarrow C = 1/2$

(i) Mean $\mu = \int_{-\infty}^{\infty} xf(x)dx. = \frac{1}{2}\int_{-\infty}^{\infty} x.e^{-|x|}dx = \infty$

= 0 [integrand is odd]

(ii) $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$

=2

(iii) $P(0 \le X \le 4)$=0.49

$\int_{0}^{4} f(x)dx = 0.49$

Assignment-Cum-Tutorial Questions

Section - A

1. Given that $P(A)=0.9,\ P(B)=0.89,\ P(A\cap B)=0.84,$ then $P(A\cup B)$ is
   (a) 0.95 (b) 0.59 (c) 0.99 (d) 0.095
2. An experiment yields three mutually exclusive events A,B,C with $P(A)=2P(B)=3P(C)$ then $P(A)$ is
   (a) 2/11 (b) 3/11 (c) 6/11 (d) 5/11
3. The probability of solving a problem by the three students A, B, C respectively are 1/3, 1/4, 1/5. Then the probability that the problem will be solved is
   (a) 1/5 (b) 2/5 (c) 3/5 (d) none
4. If two balls are drawn from a bag containing 3 white 4 black and 5 red balls, then the probability that the balls drawn are of different colours is
   (a) 47/66 (b) 10/33 (c) 5/22 (d) 2/11
5. A and $\overline{B}$ are two independent events such that $P(\overline{A}\cap B)=\frac{8}{25}$ and $P(A\cap\overline{B})=\frac{3}{25}$, then $P(A)$ is
   (a) 2/5 (b) 4/5 (c) 1/5 (d) 3/5
6. If $p(x)=x+\frac{2}{k}, x=1,2,3,4,5$ is the probability distribution of a discrete random variable, then $k=$
   (a) 5/7 (b) -5/7 (c) 7/5 (d) -7/5
7. If $f(x)=\frac{k}{(1+x^2)}, -\infty < x < \infty$ is a valid density function, then $k=$
   (a) $1/\pi$ (b) π (c) $-1/\pi$ (d) none
8. If X is a continuous random variable with probability density function
   $f(x)=\frac{(x+1)}{8},\ for\ 2<x<4$
   $=0, otherwise$
   Then E(X) =
   (a) 3.308 (b) 3.803 (c) 3.083 (d) 3.380
9. If X is a random variable and V(X) = 2, then V(2X + 3) =
   (a) 2 (b) 3 (c) 6 (d) 8
10. The relation between probability density function and cumulative density function of a random variable X is
    (a) $F(x)=\int_{-\infty}^{x} f(x)dx$ (b) $F(x)=\int_{x}^{\infty} f(x)dx$ (c) $F(x)=\int_{-\infty}^{0} f(x)dx$ (d) $F(x)=\int_{0}^{\infty} f(x)dx$
11. If $f(x)=2e^{-2x}, x>0$ is a probability density function, then P(X ≥ 0.5) =
    (a) e^{-1} (b) e^{-2} (c) e^{-3} (d) e

Section – B

1. If we draw a card from a pack, what is the probability that the card is either ace or king?
2. A die is thrown twice. What is the probability that the sum of the spots on the die at two throws is divisible by 2 or 3?

3. A bag contains 8 white and 4 red balls. One ball is drawn from the bag and it is replaced after noting its colour. In the second draw again one ball is drawn and its color is noted. What is the probability of the event that both the balls drawn are of different colours?

4. A bag contains 8 white and 4 red balls. One ball is drawn from the bag and it is not replaced after noting its colour. In the second draw again one ball is drawn and its colour is noted. What is the probability of the event that both the balls selected at two successive draws are of different colours?

5. A lot of 100 semiconductor chips have 20 defective chips. Two chips are selected at random without replacement from the lot.
   a) What is the probability that the first one selected is defective?
   b) What is the probability that the second one selected is defective, given that the first one was defective?
   c) What is the probability that both are defective?

6. If A and B are mutually exclusive events, P(A) = 0.23, and P(B) = 0.51, find
   (i) $P(\overline{A})$ (ii) $P(A \cup B)$ (iii) $P(\overline{A} \cap B)$ (iv) $P(\overline{A} \cap \overline{B})$

7. Given P(A) =0.35, P(B) = 0.73, and $P(A \cap B) = 0.14$, find
   (i) $P(A \cup B)$ (ii) $P(A \cap \overline{B})$ (iii) $P(\overline{A} \cup \overline{B})$ (iv) $P(\overline{A} \cap B)$

8. A shipment of components consists of three identical boxes. One box contains 2000 components of which 25% are defective, the second box has 5000 components of which 20% are defective and the 3rd box contains 2000 components of which 600 are defective. A box is selected at random and a component is removed at random from the box.
   (i) What is the probability that this component is defective?
   (ii) What is the probability that the defective component came from the second box?

9. Three machines A, B and C produce 55%, 25%, 20% of the total number of items of a factory. The percentage of defective output of these machines is 3%, 2% and 4%. If an item is selected at random, (i) find the probability that the item is defective (ii) if the selected item is defective, find the probability that the item is produced by machine A, machine B and machine C.

10. A random variable X has the following probability function value of X

X	0	1	2	3	4	5	6
P(X)	k	3k	5k	7k	9k	11k	13k

Find (i) k (ii) P(X < 4) (iii) P (x ≥ 5) (iv) P (X ≤ x) > ½ ?

11. Find the mean and variance of the uniform probability distribution given by f(x)=1 / n for x=1,2,…..n

12. A continuous random variable X has a pdf $f(x) = 4x^3$, for $0 \le x \le 1$. Find the values of a and b such that (i) $P(X \le a) = P(X > a)$ (ii) $P(X > b) = 0.1$. Also find the mean and variance of the random variable X.

13. Probability density function of a random variable X is

$f(x) = \frac{\sin x}{2}, \ 0 < x < \pi$
$= 0, \ elsewhere$

. Find the mean and variance of the distribution. Also calculate the probability of X lies between 0 and π/2.

14. The daily consumption of electric power (in million of KW-hours) is a random variable having the probability density function

$f(x) = \frac{1}{9} x e^{-x/3}, x > 0$
$= 0, \ x \ge 0$

.

If the total production is 12 million KW-hours, determine the probability that there is power cut (shortage) on any given day. Also find the average daily consumption of electric power.

15. A two-faced fair coin has its faced designated as head (H) and tail(T). This coin is tossed three times in succession to record the following outcomes. H, H, H. If the coin is tossed one more time. the probability (up to one decimal place) of obtaining H again, given the previous realizations of H, H and H would be _____

Answer : 0.5 **GATE- 17**

16. Probability density function of a random variable *X* is given below

f(x) = $\begin{cases} 0.25 & 1 \le x \le 5 \\ 0 & otherwise \end{cases}$ then P (x ≤4) = _____

Answer : 0.75 **GATE- 16**

17. Consider the following probability mass function (p.m.f) of a random variable X.

$p(x,q) = \begin{cases} q & if \ X = 0 \\ 1-q & if X = 1 \\ 0 & otherwise \end{cases}$

If q = 0.4, the variance of X is _____

18. The probability density function of a random variable, x is

$f(x) = \frac{x}{4}\left(4 - x^2\right) for 0 \le x \le 2$
$= 0 \ otherwise$

The mean, μ_x of the random variable is _____

UNIT II – Standard Probability Distributions

Objectives:

(a) Understand discrete probability distributions (Binomial and Poisson distributions)

(b) Understand continuous probability distribution (Normal distribution) and apply it to determine the probabilities in real world problems

Syllabus :

Discrete distributions: Binomial distribution-probability-mean, variance - Poisson distribution-probability-mean, variance - fitting of Poisson distribution. Continuous distributions: normal distribution-properties – Problems.

Learning Outcomes:

Students will be able to

(a) distinguish between discrete probability distribution and continuous probability distribution

(b) find the exact probability for X successes in n trials of a binomial experiment

(c) know where Poisson distribution can be applied.

(d) compute the mean and variance of a discrete probability distributions

(e) find the probabilities associated with a normal probability distribution using the standard normal table.

Learning Material

There are two types of probability distributions namely (1) Discrete probability distributions (Binomial and Poisson distributions) and (2) Continuous probability distributions (Normal distribution).

Binomial distribution: Binomial distribution was discovered by James Bernoulli in the year 1700 and it is a discrete probability distribution.

Where a trial or an experiment results in only two ways say 'success' or 'failure'.

Some of the situations are: (1) Tossing a coin – head or tail (2) Birth of a baby – girl or boy (3) Auditing a bill – contains an error or not.

Definition: A random variable X is said to be Binomial distribution if it assumes only non-negative values and its probability mass function is given by

$$P(X = x) = p(x) = n_{C_x}\, p^x q^{n-x},\ x = 0,1,2,.....,n$$

$$= 0,\ otherwise$$

where $q = 1 - p,\quad p + q = 1.$ Here n, p are called parameters.

Example: (1) The number of defective bolts in a box containing 'n' bolts.

(2) The number of post-graduates in a group of 'n' men.

Conditions: (1) The trials are repeated under identical conditions for a fixed number of times, say 'n' times.

(2) There are only two possible outcomes, for example success or failure for each trial.

(3) The probability of success in each trial remains constant and does not change from trail to trail.

(4) The trails are independent i.e. the probability of an event in any trail is not affected by the results of any other trail.

Constants (mean & variance) of Binomial distribution:

$$mean = E(X) = \sum xp(x)$$

$$= \sum_{x=0}^{n} x\, n_{C_x}\, p^x q^{n-x} = \sum_{x=1}^{n} x \frac{n}{x}\, n-1_{C_{x-i}}\, p^x q^{n-x} = np\sum_{x=1}^{n} n-1_{C_{x-1}}\, p^{x-1} q^{n-x} = np(q+p)^{n-1} = np$$

$$\text{var}\, iance = E(X^2) - [E(X)]^2$$

$$= \sum [x(x-1)+x]p(x) = \sum x(x-1)\,p(x) + \sum xp(x)$$

$$= \sum x(x-1)\frac{n(n-1)}{x(x-1)}\, n-2_{C_{x-2}}\, p^x q^{n-x} + np = n(n-1)\,p^2 \sum_{x=2}^{n} n-2_{C_{x-2}}\, p^{x-2} q^{n-x} + np$$

$$= n(n-1)\,p^2 (q+p)^{n-2} + np = n(n-1)\,p^2 + np = npq$$

Problem: Ten coins are thrown simultaneously. Find the probability of getting at least 7 heads.

Solution: p = probability of getting a head = ½

q = probability of getting a head = ½

The p.d.f. of binomial distribution is $P(X = x) = n_{C_x}\, p^x q^{n-x},\ x = 0,1,2,.....,10$

Given n = 10, $P(X = x) = 10_{C_x}\, p^x q^{10-x}$

Probability of getting at least 7 heads is given by

$P(X \geq 7) = P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) = 0.1719.$

Poisson distribution: Poisson distribution due to French mathematician Denis Poisson in 1837 is a discrete probability distribution.

It is a rare distribution of rare events i.e. the events whose probability of occurrence is very small but the number of trails which would lead to the occurrence of the event, are very large.

As $n \to \infty,\ p \to 0$ B.D. tends to P.D.

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

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

$$= 0,\ otherwise.$$

Here $\lambda > 0$, is called parameter of the distribution.

Example: (1) The number defective bulbs manufactured by a company.

(2) The number of telephone calls per minute at a switch board.

Conditions: (1) The variable or number of occurrences is a discrete variable.

(2) The occurrences are rare.

(3) The number of trails 'n' is large.

(4) The probability of success (p) is very small.
(5) np = λ is finite.

Constants (mean and variance) of Poisson distribution:

$$mean = E(X) = \sum_{x=0}^{\infty} x\, p(x) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda}.e^{\lambda} = \lambda.$$

$$\text{var}\, iance = E(X^2) - [E(X)]^2$$

$$= \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-1)!} + \lambda - \lambda^2 = \lambda^2 e^{-\lambda}.e^{\lambda} + \lambda - \lambda^2 = \lambda.$$

Example: Fit a Poisson distribution for the following data and calculate the expected frequencies.

X	0	1	2	3	4
f(x)	109	65	22	3	1

Solution: By the given data, total frequency = $\sum f_i$ =200

$$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i} = \frac{(0)(109)+(1)(65)+(2)(22)+(3)(3)+(4)(1)}{200} = 0.61 = \lambda$$

Therefore, the theoretical frequencies = N p(x); x = 0,1,2,3,4.

i.e. $200.\frac{e^{-0.61}(0.61)^x}{x!}$ where x = 0,1,2,3,4.

When x = 0, 200 p(0) = 108.67
x = 1, 200 p(1) = 66.29
x= 2, 200 p(2) = 20.22
x= 3, 200 p(3) = 4.11
x= 4, 200 p(4) = 0.63

since frequencies are always integers, therefore by converting them to nearest integers, we get

Observed frequency	109	65	22	3	1
Expected frequency	109	66	20	4	1

Example: A car hire firm has two cars which it hires out day by day, The number of demands for a car on each day is distributed as a Poisson distribution with mean 1.5. Calculate the proportion of days (i) on which there is no demand (ii) on which demand is refused.
Solution: Given mean, $\lambda = 1.5$

We have $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

(i) P(no demand) = p(0) = 0.2231
Number of days in a year there is no demand of car = 365 (0.2231) = 81 days.
(ii) P(demand refused) = p(x>2) = 1- [p(0)+p(1)+p(2)] = 0.1913
Number of days in a year when some demand is refused = 365 (0.1913) = 70 days.

Normal Distribution: It was first discovered by English Mathematician De-Moivre in 17333 and further refined by French Mathematician Laplace in 1744 and independently by Karl Friedrich Gauss. Normal distribution is also known as 'Gaussian distribution'.

Definition: A random variable X is said to have a Normal distribution, if its probability density function is given by

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

Where μ is mean and σ^2 is variance are called parameters.
Notation: X~N(μ, σ^2).
Problem: In a normal distribution, 7% of the items are under 35 and 89% under 63. Determine the mean and variance of the distribution.
Solution: Given P(X<35) = 0.07 and P(X<63) = 0.89

Therefore, P(X>63) = 1- P(X<63) = 1 – 0.89 = 0.11
When X = 35, Z = (X-μ)/σ = (35-μ)/σ = $-z_1$ (say) ……(1)
When X = 63, Z = (X-μ)/σ = (63-μ)/σ = $-z_2$ (say) …….(2)
P(0<Z<z_2) = 0.39 $\Rightarrow z_2$ = 1.23 (from tables)
and P(0<Z<z_1) = 0.43 $\Rightarrow z_1$ = 1.48
from(1) we have (35-μ)/σ = -1.48 ……(3)
from (2) we have (63-μ)/σ = 1.23 ……(4)
(4) – (3) gives σ = 10.332
From equation (3), μ = 50.3
Therefore, mean = 50.3 and variance = 106.75

Characteristics:
(1). The graph of the Normal distribution y = f(x) in the xy-plane is known as the normal curve.
(2). The curve is a bell shaped curve and symmetrical with respect to mean i.e., about the line x =μ and the two tails on the right and the left sides of the mean (μ) extends to infinity. The top of the bell is directly above the mean μ.
(3). Area under the normal curve represents the total population.
(4). Mean, median and mode of the distribution coincide at x = μ as the distribution is symmetrical. So normal curve is unimodal (has only one maximum point).
(5). x-axix is an asymptote to the curve.
(6). Linear combination of independent normal variates is also a normal variate.
(7). The points of inflexion of the curve are at x = μ ± σ and the curve changes from concave to convex at x = μ + σ to x = μ - σ.
(8). The probability that the normal variate X with mean μ and standard deviation σ lies between x_1 and x_2 is given by

$$P(x_1 \le X \le x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \quad \text{…….(1)}$$

Since (1) depends on the two parameters μ and σ, we get different normal curves for different values of μ and σ and it is an impracticable task to plot all such normal curves. Instead, by putting z=(x-μ)/σ, the R.H.S. of equation (1) becomes independent of the two parameters μ and σ. Here z is known as the standard variable.
(9). Area under the normal curve is distributed as follows:
P(μ - σ<X< μ + σ) = 0.6826 ; P(μ - 2σ<X< μ + 2σ) = 0.9543; P(μ - 3σ<X< μ + 3σ) = 0.9973.

Assignment-Cum-Tutorial Questions

Section - A

1. The graph of the normal curve is symmetric about the line
   (a) $x = \mu$ (b) $x = -\mu$ (c) $x = 0$ (d) $x = \pi$
2. The mean of a Poisson distribution is 8 then its variance is
   (a) 64 (b) 4 (c) 8 (d) none
3. A coin is tossed 3 times then the probability of obtaining two heads will be
   (a) 1/8 (b) 3/8 (c) 5/8 (d) 7/8
4. Mean, median and mode are equal for
   (a) Normal distribution (b) Binomial distribution (c) Poisson distribution (d) Bernoulli distribution.
5. For a Poisson variate, probability of getting at least one success is
   (a) $1-e^{-\lambda}$ (b) $1-e^{\lambda}$ (c) $1+e^{-\lambda}$ (d) $1+e^{\lambda}$
6. If X is a Poisson random variable such that $2P(X = 0) = P(X = 2)$ then the standard deviation of X is
   (a) 2 (b) $\sqrt{2}$ (c) ½ (d) $1/\sqrt{2}$
7. In the standard normal curve the area between $z = -1$ and $z = 1$ is nearly
   (a) 90% (b) 95% (c) 68% (d) 75%
8. Among the items manufactured in a factory, 5% are defective. The probability of getting one defective blade in a pack of 5 blades is
   (a) 0.2044 (b) 0.4022 (c) 0.2404 (d) 0.0244
9. Mean is always greater than variance for
   (a) Normal distribution (b) Binomial distribution (c) Poisson distribution (d) none
10. Binomial distribution is used in communication systems is
    (a) inappropriate (b) false (c) true (d) none

11. The mean of Binomial distribution is _____.
12. If the mean of the binomial distribution is 6 and variance is 2, then p = _____.
13. Write the probability law of binomial distribution whose mean is 5 and variance is 10/3. _____.

Section - B

1. Consider a simple trial of tossing a fair coin six times. Calculate the probabilities of getting (i) E_1: exactly three heads, (ii) E_2: at least three heads, (iii) E_3: not more than two heads.
2. 10% of the bolts produced by a certain machine turn out to be defective. Find the probability that in a sample of 10 bolts selected at random exactly two will be defective using (i) Binomial distribution (ii) Poisson distribution and comment on the results.

3. If a bank receives on an average 6 bad checks per day. What are the probabilities that it will receive (i) four bad checks on any given day? (ii) 10 bad checks over any two consecutive days?
4. A safety engineer feels that 30% of all industrial accidents in his plant are caused by failure of employees to follow instructions. If this figure is correct, find approximately, the probability that among 84 industrial accidents in the plant, anywhere from 20 to 30 (inclusive) will be due to the failure of employees to follow instructions.
5. If the probability that an individual suffers a bad reaction due to a certain injection is 0.001, determine the probability that out of 2000 individuals (i) exactly 3 (ii) more than 2 individuals will suffer a bad reaction.
6. The number of calls arriving on an internal switch board of an office is 90 per hour. Calculate the probability of 1 to 3 calls in a minute on the switch board.
7. The number of mistakes counted in one hundred typed pages of a typist revealed that he made 2.8 mistakes on an average per page. Find the probability that (i) there is no mistake (ii) there are two or less mistakes in a page typed by him.
8. In a test on 1000 electric bulbs, it was found that the number of bulbs was normally distributed with an average life of 2040 hours and a standard deviation of 60 hours. How many bulbs are likely to be in usage for (a) more than 2150 hours (b) less than 1950 hours (c) more than 1920 hours but less than 2100 hours.
9. Life time of IC chips manufactured by a semiconductor manufacturer is approximately normally distributed with mean 5×10^6 hours and standard deviation of 5×10^5 hours. A mainframe manufacturer requires that at least 95% of a batch should have a lifetime greater than 4×10^6 hours. Will the deal be made?
10. Find the probabilities that a random variable having the standard normal distribution will take a value (i) between 0.87 and 1.28 (ii) between -0.34 and 0.62 (iii) greater than 0.85 (iv) greater than -0.655 along with neat diagrammatic representation.

11. In a certain junior Olympics, javelin throw distances are well approximated by a Gaussian distribution for which $\mu = 30$m and $\sigma = 5$m. In a qualifying round, contestants must throw farther than 26m to qualify. In the main event, the record throw is 42m.
    (i) What is the probability of being disqualified in the qualifying round?
    (ii) In the main event, what is the probability that the record will be beaten?

12. Fit a Poisson distribution to the following data.

x	0	1	2	3	4
f(x)	109	65	22	3	1

PROBABILITY & STATISTICS
UNIT-III
SAMPLING DISTRIBUTIONS

Objectives:

- To know Sampling Theory.
- Understand properties of Point Estimators

Syllabus:

Population and sample-types of sampling-Sampling distribution of mean - Sampling distribution of sums and differences. Point and interval estimation, Confidence Interval for mean and proportions.

Learning Outcomes:

Students will be able to

- Understand the concept of a sampling distribution
- Describe the distribution of the sample mean for samples obtained from a population that is not normal
- Understand Point Estimation and be able to compute point estimates

Learning Material
UNIT-3 : SAMPLING DISTRIBUTIONS

Basic Terms:
Population: In statistics population does not only refers to people but it may defined as any collection of individuals or objects or units which can be specified numerically.
Population may be mainly classified into two types.
(i) Finite population (ii) Infinite population
(i) Finite population: The population contains finite number of individuals is called 'finite population'. For example, total number of students in a class.
(ii) Infinite population: The population which contains infinite number of individuals is known as 'infinite population'. For example, the number of stars in the sky.
Parameter: The statistical constants of a population are known as parameter.
For example, mean (μ) and variance (σ^2).
Statistic: Any function of sample observations is called sample statistic or statistic.
Standard error: The standard deviation of the sampling distribution of a statistic is known as its 'standard error'.

Sample: A portion of the population which is examined with a view to determining the population characteristics is called a sample. Or A sample is a subset of the population and the number of objects in the sample is called the size of the sample size of the sample is denoted by 'n'.

Classification of samples: Samples are classified in 2 ways.

(i) **Large sample:** The size of the sample (n) $\geq$ 30, the sample is said to be large sample.

(ii) **Small sample:** If the size of the sample (n) < 30, the sample is said to be small sample or exact sample.

Types of sampling: There are mainly 5 types of sampling as follows.

(i) **Purposive sampling:** It is one in which the sample units are selected with definite purpose in view. For example, if you want to give the picture that the standard of living has increased in the town of 'Gudivada', we may take individuals in the sample from Satyanarayana puram, Rajendra nagar etc and ignore the localities where low income group and middle class families live.

(ii) **Random sampling:** In this case sample units are selected in one in which each unit of population has an equal chance of being included in it.
Suppose we take a sample of size 'n' from finite population of size 'N'. Then there are N_{Cn} possible samples. A sampling technique in which each of the N_{Cn} samples has an equal chance of being selected is known as 'Random sampling' and the sample obtained by this is termed as 'random sample'.

(iii) **Stratified Random Sampling:** It is defined as the entire heterogeneous population is sub divided into homogeneous groups. Such groups are called 'strata'. The size of each strata may differ but they are homogeneous within themselves. A sample is drawn randomly from these strata's is known as 'stratified random sampling'.

(iv) **Systematic sampling:** In this sampling we select a random number to draw a sample and the remaining samples are automatically selected by a predetermined patterns such a sampling is called 'systematic sampling'.

(v) **Simple sampling:** Simple sampling is random sampling in which each unit of the population has an equal chance. For example, if the population consists of N units then we select a sample n units then each unit having equal probability 1/N.

Problems:

(1) Find the value of the finite population correction factor for n = 10 and N = 1000.
Given N = the size of the finite population = 1000
n = size of the sample = 10
Therefore, correction factor = (N-n)/(N-1) = 0.991

(2) A population consists of five numbers 2, 3, 6, 8 and 11. Consider all possible samples of size two which can be drawn with replacement from this population. Find
(a) The mean of the population (b) standard deviation of the population (c) mean of the sampling distribution of means (d) standard deviation of the sampling distribution of means.

Solution: (a) Mean of the population
$\mu = (2+3+6+8+11)/5 = 6$

(b) Variance (σ^2) is $\sigma^2 = \sum (x_i - \bar{x})^2 / n$

$= (2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2 / 5 = 10.8$

Therefore, standard deviation (s.d.) $\sigma = \sqrt{10.8} = 3.29$

(c) Sampling with replacement:

Total number of samples with replacement is $N^n = 5^2 = 25$ samples of size 2. i.e.
{(2,2), (2,3),(2,6),(2,8),(2,11),(3,2), (3,3),(3,6),(3,8),(3,11),(6,2),(6,3),(6,6),(6,8),(6,11)
(8,2), (8,3),(8,6),(8,8),(8,11),(11,2),(11,3),(11,6),(11,8),(11,11)}

Therefore, the distributin of means of the samples known as sampling distribution of means.

Therefore, the samples means are {2, 2.5, 4, 5, 6.5, 2.5, 3, 4.5, 5.5, 7, 4, 4.5, 6, 7, 8.5, 5, 5.5, 7, 8, 9.5, 6.5, 7, 8.5, 9.5, 11} and the mean of sampling distribution of means is the mean of these 25 means.

$\mu_x = (2+2.5+4+ \ldots +9.5 +11)/25 = 6.$

(d) Standard deviation:

$\sigma^2 = (2-6)^2+(2.5-6)^2+ \ldots +(9.5-6)^2+(11-6)^2 / 25 = 5.40$

Therefore , $\sigma = \sqrt{5.40} = 2.32$

(3) A population consists of 5, 10, 14, 18, 13, 24. Consider all possible samples of size 2 which can be drawn without replacement from the population. Find

(a) The mean of the population (b) standard deviation of the population (c) mean of the sampling distribution of means (d) standard deviation of the sampling distribution of means.

Solution: (a) Mean of the population

$$\mu = \frac{\sum x}{n} = (5+10+14+18+13+24)/6 = 14$$

(b) Variance (σ^2) is $\sigma^2 = \sum (x_i - \bar{x})^2 / n$

$$= (5-14)^2+(10-14)^2+\ldots+(11-6)^2 / 6 = 35.67$$

Therefore, standard deviation (s.d.) $\sigma = \sqrt{10.8} = 3.29$

(c) All possible samples of size 2 i.e. the number os samples = $16_{C2} = 15$

Sample No.	Sample Values	Total of Sample values	Sample mean
1	5, 10	15	7.5
2	5, 14	19	9.5
3	5, 18	23	11.5
4	5, 13	18	9
5	5, 24	29	14.5
6	10, 14	24	12
7	10, 18	28	14
8	10, 13	23	11.5
9	10, 24	34	17
10	14,18	32	16
11	14, 13	27	13.5
12	14, 24	38	19
13	18, 13	31	15.5
14	18, 24	42	21
15	13, 24	37	18.5
		Total	210.0

(d) Variance of sampling distribution of means

$$\sigma_{\bar{X}}^2 = \frac{(7.5-14)^2 + (9.5-14)^2 + ... + (21-14)^2 + (18.5-14)^2}{15} = 14.266$$

Therefore, standard deviation, $\sigma_{\bar{X}} = \sqrt{14.266} = 3.78$

Sampling distribution of differences and sums:

(1) Let $u_1 = (3,7,8)$, $u_2 = (2,4)$. Find (a) μ_{u_1} (b) μ_{u_2} (c) mean of the sampling distribution of the difference of means $\mu_{u_1-u_2}$ (d) the s.d of the sampling distribution of the differences of means ($\sigma_{u_1-u_2}$).

Solution: Given $u_1 = (3,7,8)$, $u_2 = (2,4)$

$u_1 - u_2 = \{1,-1,5,3,6,4\}$

(a) μ_{u_1} =(3+7+8)/2 = 6 (b) μ_{u_2} = (2+4)/2 = 3 (c) $\mu_{u_1-u_2}$ = (1+5+3+6+4-1)/6 = 3

(d) $\sigma_{u_1} = \sqrt{(6-3)^2 + (6-7)^2 + (6-8)^2/3} = \sqrt{14/3}$ (d) $\sigma_{u_2} = \sqrt{(2-3)^2 + (3-4)^2/2} = 1$

(e) $\sigma_{u_1-u_2} = \sqrt{(1-3)^2 + (5-3)^2 + ... + (3-3)^2 + (-1-3)^2/6} = \sqrt{17/3}$.

Note:

(1) If $X \sim N(\mu, \sigma^2)$ then $Z = \frac{X-\mu}{\sigma}$.

(2) If $\bar{X} \sim N(\mu, \sigma^2/n)$ then $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.

Problem: The mean of certain normal population is equal to the standard error of the mean of the samples of 64 from that distribution. Find the probability that the mean of the sample size 36 will be negative.

Solution: Given sample size n =64

and mean μ = standard error of the mean of the samples

we have standard error of mean = $\sigma/\sqrt{n}$

$\therefore$ mean $\mu = \sigma/\sqrt{64} = \sigma/8$

We know that $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = = \frac{\bar{X}-\sigma/8}{\sigma/\sqrt{36}} = (6\bar{X}/\sigma) - (3/4).$

If Z < 0.75, $\bar{X}$ is negative

$$P(Z < 0.75) = P(-\infty < Z < 0.75) = \int_{-\infty}^{0} \phi(z)dz + \int_{0}^{0.75} \phi(z)dz = 0.5 + 0.2734 = 0.7734.$$

Problem: A random sample of size 100 is taken from an infinite population having the mean μ = 76 and the variance σ^2 =256. What is the probability that $\bar{x}$ will be between 75 and 78.

Solution: Given n= 100, μ = 76 and the variance σ^2 =256.

We have $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

When $\bar{x}$ =75, $z_1 = \frac{75-76}{16/\sqrt{100}} = -0.625$

$$\bar{x} = 78,\ z_2 = \frac{78-76}{16/\sqrt{100}} = 1.25$$

$$\therefore P(75 < \bar{x} < 78) = P(z_1 \le Z \le z_2) = P(-0.625 \le Z \le 1.25) = P(-0.625 \le Z \le 0) + P(0 \le Z \le 1.25)$$

$$= 0.2334 + 0.3944 = 0.628.$$

Note:

(1) If $X \sim N(\mu, \sigma^2)$ then $Z = \frac{X-\mu}{\sigma}$.

(2) If $\bar{X} \sim N(\mu, \sigma^2/n)$ then $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.

ESTIMATION:

Introduction: Theory of estimation was founded by Prof. R.A.Fisher in a series of fundamentals in 1930's. The statistic whose distribution concentrates as closely as possible near the true value of the parameters may be regarded as the best estimate i.e. determining the functions of sample observations $\theta_1(x_1, x_2, ..., x_n)$, $\theta_2(x_1, x_2, ..., x_n), ..., \theta_n(x_1, x_2, ..., x_n)$ such that their distribution is concentrated as closely possible near the true value of the parameter.

Definitions:

Estimator: The statistic $\hat{\theta}_n$ which is used to estimate unknown population parameter is known as 'estimator'.

Estimate: The value of 'estimator' is known as 'estimate'.

Estimation: The statistical technique of estimating an unknown parameters from the corresponding sample statistics is known as 'estimation'.

There are two kinds of estimates to determine the statistic of the population parameter namely

(i) Point estimation (ii) Interval estimation

Point estimation: A particular value of statistic which is used to estimate a given parameter is known as point estimation or estimator of the parameter.

Interval estimation: An interval estimate refers to the range of values used to estimate an unknown population parameter together with the probability or confidence level that the interval does include the unknown population parameter.

Properties of good estimator or criteria of good estimator:

A good estimator is one which is very closer to the true value or actual value of the parameter as possible. Any estimator is said to be good estimator if it follows the following properties.

(i) Unbaisedness (ii) Consistency (iii) Efficiency (iv) Sufficiency

Unbiasedness: A statistic $\hat{\theta}$ is said to be an unbiased estimator of θ if $E(\hat{\theta}_n) = \theta$ for all θ.

For example, sample mean, $\bar{x}$ is an unbiased estimator of population mean, μ.

Consistency: an estimator $\hat{\theta}_n$ of a parameter θ is consistent if it converges to θ, as $n \to \infty$.

For example, sample mean, $\bar{x}$ is an consistency estimator of population mean, μ.

Efficiency: A statistic $\hat{\theta}_1$ is said to be more efficient unbiased estimator of the parameter θ than the statistic $\hat{\theta}_2$ if

(a) $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators of θ.

(b) $V(\hat{\theta}_1) < V(\hat{\theta}_2)$.

Sufficiency: An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter.

Confidence interval estimates of parameters:

The unknown parameter θ is included in the interval $[t_1, t_2]$ in a specified percentage of cases, then the interval is called a confidence interval for the parameter θ.

Confidence limits for population mean:

(i) 95% confidence limits are $\bar{x} \pm 1.96(S.E.of\ \bar{x})$

(ii) 98% confidence limits are $\bar{x} \pm 2.58(S.E.of\ \bar{x})$

(iii)99.73% confidence limits are $\bar{x} \pm 3(S.E.of\ \bar{x})$

(iv)90% confidence limits are $\bar{x} \pm 1.64(S.E.of\ \bar{x})$

Confidence limits for population proportion, P:

(i) 95% confidence limits are $p \pm 1.96(S.E.of\ p)$

(ii) 99% confidence limits are $p \pm 2.58(S.E.of\ p)$

(iii)99.73% confidence limits are $p \pm 3(S.E.of\ p)$

(iv)90% confidence limits are $p \pm 1.64(S.E.of\ p)$

Confidence limits for the difference $\mu_1 - \mu_2$ of two population means μ_1 and μ_2:

(i) 95% confidence limits are $(\bar{x}_1 - \bar{x}_2) \pm 1.96(S.E.of\ (\bar{x}_1 - \bar{x}_2))$

(ii) 99% confidence limits are $(\bar{x}_1 - \bar{x}_2) \pm 2.58(S.E.of\ (\bar{x}_1 - \bar{x}_2))$

(iii)99.73% confidence limits are $(\bar{x}_1 - \bar{x}_2) \pm 3(S.E.of\ (\bar{x}_1 - \bar{x}_2))$

(iv)90% confidence limits are $(\bar{x}_1 - \bar{x}_2) \pm 1.64(S.E.of\ (\bar{x}_1 - \bar{x}_2))$

Confidence limits for the difference $P_1 - P_2$ of two population means P_1 and P_2:

(i) 95% confidence limits are $(p_1 - p_2) \pm 1.96(S.E.of\ (p_1 - p_2))$

(ii) 99% confidence limits are $(p_1 - p_2) \pm 2.58(S.E.of\ (p_1 - p_2))$

(iii)99.73% confidence limits are $(p_1 - p_2) \pm 3(S.E.of\ (p_1 - p_2))$

(iv)90% confidence limits are $(p_1 - p_2) \pm 1.64(S.E.of\ (p_1 - p_2))$

Confidence interval for μ, σ known:

If $\bar{x}$ is the mean of a random sample of size n from the population with known variance σ^2, (1-α)100%. Confidence interval for μ is given by $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, where $z_{\alpha/2}$ is the z-value leaving an area of α/2 to the right.

Therefore, the maximum error of estimate E with (1-α) peobability is given by E = $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Confidence interval for μ, σ unknown:

If $\bar{x}$ and S are the mean and standard deviation of a random sample from a normal population with unknown variance σ^2, a(1- α)100% confidence interval for μ is $\bar{x} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$, where $t_{\alpha/2}$ is the t-value with $\upsilon = (n-1)$ d.f., leaving an area of α/2.

Maximum error (E_{max}) = $t_{\alpha/2}\frac{S}{\sqrt{n}}$.

Problems:

(1) What is the maximum error one can expect to make with probability 0.90 when using the mean of a random sample of size n = 64 to estimate the mean of population with $\sigma^2 = 2.56$.

Solution: Given n = 64, $\sigma^2 = 2.56 \Rightarrow \sigma = \sqrt{2.56} = 1.6$

And the probability = 0.90 then confidence limit = 90%

Therefore, $z_{\alpha/2} = 1.645$

Therefore, maximum error $E_{max} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.645\ (1.6/\sqrt{64}) = 0.329$.

(2) Assuming that σ = 20.0, how large a random sample be take to assert with probability 0.95 that the sample mean will not differ from the true mean by more than 3.0 points?

Solution: Given σ = 20

Probability = 0.95 = 95%

$z_{\alpha/2} = 1.96$ and E = 3 points

We know that $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$$\Rightarrow \sqrt{n} = \frac{z_{\alpha/2}\sigma}{E} \Rightarrow n = \left(z_{\alpha/2}\sigma/E\right)^2$$

$$\Rightarrow n = \left(\frac{1.96\ x 20}{3}\right)^2 = 171 (approximately)$$

(3) Find 95% confidence limits for the mean of a normality distributed population from which the following samples was taken 15, 17, 10, 18, 16, 9, 7, 11, 13, 14.

Solution: we have $\bar{x}$ = (15+ 17+ 10+ 18+ 16+ 9+ 7+ 11+ 13+ 14)/10 = 13.

Variance, $S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$

$$= \frac{1}{9}\left[(15-13)^2 + (17-13)^2 + ... + (14-13)^2\right] = 40/3$$

and 95% confidence limits are $t_{\alpha/2}$=1.96

$$\Rightarrow t_{\alpha/2}\frac{S}{\sqrt{n}} = 1.96\frac{\sqrt{40}}{\sqrt{3}\sqrt{10}} = 2.26$$

Therefore, confidence limits are $\bar{x} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$ = 13 ± 2.26 = (10.74, 15.26).

(4) A random sample of 100 teachers in a large metropolitan area revealed a mean weekly salary of Rs. 487 with a standard deviation Rs. 48. With what degree of confidence can we assert that the average weekly salary of all teachers in the metropolitan area is between 472 to 502?

Solution: Given μ = 487, σ = 48, n = 100

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 487}{48/\sqrt{100}} = \frac{\bar{x} - 487}{4.8}$$

When $\bar{x} = 472 \Rightarrow z_1 = \frac{472-487}{4.8} = -3.125$

When $\bar{x} = 502 \Rightarrow z_2 = \frac{502-487}{4.8} = 3.125$

Let X be the mean salary of teacher. Then

$$P(472 < X < 502) = P(-3.125 < Z < 3.125) = 2(0 < Z < 3.125) = 2\int_0^{3.125} \phi(z)dz = 2(0.4991) = 0.9982$$

Therefore, we can ascertain with 99.82 % confidence.

Assignment-cum-Tutorial Questions
SECTION-A

1. The number of possible samples of size n for a population of N units with replacement is _____.
2. The number of possible samples of size n for a population of N units without replacement is _____.
3. Sample variance formula is _____.
4. The difference between sample estimate and population parameter is called _____.

5. 100 among 600 articles are defective. If the maximum error with probability 0.99 is 0.02. The sample size is _____.
6. If there are 5 defective items among 4000, one sided 99% confidence interval for proportion is _____.
7. If n = 144, σ = 4, $\bar{x}$ = 150 then 95% confidence interval for μ is _____.
8. If the maximum error with probability 0.95 is 1.2, and standard deviation of the population 10. Then sample size is _____.
9. A sample size 100 is taken whose standard deviation is 5. What is the maximum error with probability 0.95 _____.
10. The totality of the observation called
    (a) Population (b) Sample (c) Parameter (d) None
11. The statistical constants of the population are called
    (a) Statistic (b) Parameter (c) Sample Statistic (d) One
12. The finite population correction factor is
    (a) $\frac{n-N}{N-1}$ (b) $\frac{N-n}{N-1}$ (c) $\frac{N-1}{N-n}$ (d) None
13. The standard error of the statistic sample mean $\left(\overline{X}\right)$is
    (a) $\frac{\sigma}{\sqrt{n}}$ (b) $\frac{\sigma^2}{\sqrt{n}}$ (c) $\sqrt{\frac{\sigma}{n}}$ (d) None
14. If $\overline{X}$ is the mean of a random sample of size n from a finite population of size N with the mean μ and the variance σ^2 then
    (a) $\mu\frac{\sigma^2}{n}$ (b) $\mu,\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ (c) μ,σ (d) None
15. If $\overline{X}$ =157, $\mu = 155$, $\sigma = 15$ and n = 36 then z is
    (a) 0.8 (b) 0.6 (c) 0.08 (d) None
16. If n = 40, $\sigma = 2.06$ then the maximum error with 99% confidence is
    (a) 0.7377 (b) 0.8387 (c) 0.6387 (d) 0.536
17. A sample of size n is taken from a population whose variance is 9. The maximum error of estimate for μ with 95% confidence is 0.5. Then n =
    (a) 12 (b) 68 (c) 128 (d) 139
18. If n = 144, $\sigma = 4$ *and* $\bar{x} = 32$ then 99% confidence interval for means is
    (a) (30.71,33.29) (b) (30.835,33.165 (c) (31.02,32.98) (d) None

19. If the maximum error with 99% confidence is 0.86 and size of the sample is 144, then the variance of the population is

    (a) 2 (b) 4 (c) 8 (d) 16

20. If the size of the sample is 5 and size of the population is 2000. The correction factor is

    (a) 0.999 (b) 9.99 (c) 99.9 (d) None

SECTION-B

(II) Descriptive Questions:

1. A population consists of five numbers 2, 3, 6, 8 and 11. Consider all possible samples of size two which can be drawn without replacement from this population. Find the mean of the population (b) standard deviation of the population (c) mean of the sampling distribution of means (d) standard deviation of the sampling distribution of means.
2. A population consists of 5, 10, 14, 18, 13, 24. Consider all possible samples of size 2 which can be drawn without replacement from the population. Find the mean of the population (b) standard deviation of the population (c) mean of the sampling distribution of means (d) standard deviation of the sampling distribution of means.
3. $U_1 = \{5,6,7,8\}$ $u_2 = \{10,12,14\}$ write (i) $u_1 + u_2$ (ii)) $u_1 - u_2$ (iii) $\mu_{u_1+u_2}$ (iv) $\mu_{u_1-u_2}$
4. Assume that the heights of 3000 male students at a college are normally distributed with mean 68 inches and standard deviation 3 inches. If 80 samples consisting of 25 students each are obtained, what would be the expected mean and standard deviation of the resulting sampling distribution of means if the sampling were done (a) with replacement (b) without replacement.
5. Determine the expected number of random samples having their means (a) between 22.39 and 22.41 (b) greater than 22.42 (c) less than 22.37 (d) less than 22.38 or more than 22.41 for the following data: N = 1500, n = 36, number of samples = 300, $\mu = 22.4$, $\sigma = 0.48$.
6. A certain type of electric light bulb has a mean life-time of 1500h and a standard deviation of 150h. Three bulbs are connected so that when one burns out, another will go on. Assuming that the life-time are normally distributed, what is the probability that lighting will take place for (a) at least 5000h and (b) at most 4200h?
7. Determine the probability that the mean breaking strength of cables produced by company 2 will be (i) at least 600N more than (ii) at least 450N more than the cables produced by company 1, if 100 cables of brand 1 and 50 cables of brand 2 are tested.
8. The mean voltage of a battery is 15 volt and s.d. is 0.2 volt. What is the probability that four such batteries connected in series will have a combined voltage of 60.8 or more volts?
9. In a random sample, 136 of 400 persons given a flu vaccine experienced some discomfort. Construct a 95% confidence interval for the true proportion of persons who will experience some discomfort from the vaccine.
10. A district official intends to use the mean of a random sample of 150 sixth grades from a very large school district to estimate the mean score which all the sixth grades in the district would get if they took a certain arithmetic achievement test. If based on experience, the official known that $\sigma = 9.4$ for such data, what can she assert with probability 0.95 about the maximum error?

11. The mean of certain normal population is equal to the standard error of the mean of samples of size 64. Find the probability that the mean of the sample size 36 will be negative.

Unit-IV : Testing Of Hypothesis (Large samples)

Objectives:

- Understand how to develop Null and Alternative Hypotheses
- Know the principles of hypothesis testing

Syllabus:

Null hypothesis-Alternative hypothesis-level of significance-degrees of freedom. Type I and Type II errors- One tail and two tailed tests - Testing of hypothesis concerning means and proportions

Learning Outcomes: The students will be

- Able to setup null and alternative hypothesis
- Able to do hypothesis test about population mean
- Able to do hypothesis test about population proportion.
- apply a range of statistical tests appropriately.

Statistical Hypothesis: Hypothesis is a statement or assumption about the population which may or may not be true

Testing of hypothesis: It is used to testing the hypothesis about the parent population from which the samples are drawn.

Test of Significance: A very important aspect of the sampling theory is the study of the test of significance, which enables us to decide on the basis of the sample results, if

- The deviation between the observed sample statistics and the hypothesis parameter value (or)
- The deviation between two independent sample statistics is significant.

Null Hypothesis: A definite statement about the population parameter. Such hypothesis which is usually a hypothesis of no difference is called 'Null hypothesis' and is usually denoted by 'H_0'.

Alternative Hypothesis: Any hypothesis which is complementary to the null hypothesis is called 'Alternative hypothesis" and is usually denoted by 'H_1'.

Eg: If we want to test the null hypothesis that the population has a specified mean μ_0 (say) i.e., H_1:$\mu \neq \mu_0$----------(i)

H_1:$\mu < \mu_0$----------(ii)

H_1:$\mu > \mu_0$----------(iii)

Then the alternative hypothesis in (i) is known as Two-tailed test and the alternatives in (ii)and (iii) are known as left and right tailed tests respectively.

Critical Region: The region of rejection of null hypothesis H_0 when H_0 is true is that region of the outcomes at where H_0 is rejected. If the sample points falls in that region is called the 'critical region', size of the critical region is α.

Type-I error:

P(rejecting H_0 /H_0 is true) i.e., when H_0 is true it is to be accepted but it is a rejected. Therefore there is a error

Type-II error: P(Accepting H_0/H_0 is false) i.e., when H_0 is false it is to be rejected but it is accepted. Therefore there is a error

One tailed and two tailed tests: If the alternative hypothesis is of the type (< or >) and the entire critical region lies in the normal probability curve on one side then it is said to be one tailed tests (OTT)

Again the one tailed test is two types (i) Right one tailed test (ii) Left one tailed test

If the alternative hypothesis is of the type ($\neq$) and the

critical region lies in the normal probability curve on both sides then it is said to be two tailed tests (TTT)

Level of significance (LOS): The probability of committing Type-I error is known as the level of significance which is denoted by 'α'. Usually LOS are 10% , 5% or 1%.

Degrees of freedom: Suppose there are N observations and k conditions on these then the degrees of freedom is N-k. The degrees of freedom is used in small sample tests.

Procedure for testing of hypothesis:

Step (1): Set up Null hypothesis (H_0)

Step (2): Set up Alternative hypothesis (H_1) Which enables us to apply one tailed test/ Two tailed test.

Step (3): Choose Level of significance (LOS) α

Step (4): Under the null hypothesis H_0, the test statistic Z=$\frac{t-E(t)}{S.E\ of(t)} \sim N(0,1)$ where 't ' is a statistic

Step (5): Conclusion: If calculated Z < (tabulated) Z_α at α % LOS then accept null hypothesis otherwise reject null hypothesis.

The rejection rule for H_0: x=μ (or) μ=μ_0 is given below.

Table: Critical value of Z when n≥30			
Level of significance	1%	5%	10%
Two-Tailed test	2.58	1.96	1.645
Right-Tailed test	2.33	1.645	1.28
Left-Tailed test	-2.33	-1.645	-1.28

Test of significance for a single mean: Working Rule

Step (1): Null hypothesis:H_0:μ=μ_0

Step (2): Alternative hypothesis: H_1:μ≠ μ_0/ H_1:μ<μ_0 / H_1:μ>μ_0

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: We have the following two cases.

Case (1): When the S.D (σ) of population is known. Then the test statistic is

$$Z=\frac{\overline{X}-\mu}{S.E(\overline{X})}\sim N(0,1)$$

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

Where σ = standard deviation of population

n=Sample size.

Case (2): When the S.D (σ) of population is unknown. The test statistic is $Z=\frac{\overline{X}-\mu}{S.E(\overline{X})}$

Where S.E ($\overline{x}$) =$s/\sqrt{n}$

Where s = standard deviation of sample

n=Sample size.

Step (5): Conclusion: Z_{cal} is compare with Z_{tab} value.

If Z_{cal}<Z_{tab} accept H_0. Otherwise reject H_0.

Problem:

A sample of 400 items is taken from a population whose standard deviation is 10. The mean of the sample is 40. Test whether the sample has come from a population with mean 38.Also calculate 95% confidence interval for the population?

Given n=400, $\bar{x}$ =40, μ=38, $\sigma = 10$

Step (1): Null hypothesis:H_0:μ=38

Step (2): Alternative hypothesis: H_1:μ≠ 38

Step (3): Level of significance: α =5%

Step (4): Test statistic: When the S.D (σ) of population is known. Then the test statistic is

Z= $\frac{\overline{X} - \mu}{S.E(\bar{X})}$ Where S.E ($\overline{x}$) =$\sigma/\sqrt{n}$

=4

Step (5): Conclusion: Z_{cal} =4, Z_{tab} =1.96

If Z_{cal}> Z_{tab} at 5% LOS. So we reject H_0.

95% confidence interval is ($\bar{x} \pm 1.96\ \sigma/\sqrt{n}$) =(39.02,40.98)

Test of equality of Two means: Let $\bar{x}_1, \bar{x}_2$ be the sample means of two independent random samples sizes n_1 and n_2 drawn from two populations having the means μ_1 and μ_2 and standard deviation σ_1 andσ_2. To test whether the two population means are equal .

Step (1): Null hypothesis: $H_0: \mu_1=\mu_2$

Step (2): Alternative hypothesis:

$$H_1 : \mu_1 \neq \mu_2 / H_1 : \mu_1 \leq_2 / H_1 : \mu_1 \geq \mu_2$$

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: Z= $\frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma^2_2}{n2}}}$

Step (5): Conclusion: Z_{cal} is compare with Z_{tab} value.

If Z_{cal}<Z_{tab} accept H_0. Otherwise reject H_0.

Problem:

The mean of two large samples of sizes 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from same population of s.d 2.5 inches?

Given n_1=1000,n_2=2000 and $\overline{x_1}$ =67.5 $\overline{x_2}$ =68 population S.D σ=2.5

Step (1): Null hypothesis:H_0:$\mu_1=\mu_2$

Step (2): Alternative hypothesis:

$$H_1: \mu_1 \neq \mu_2$$

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: Z= $\frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma^2_2}{n_2}}}$ =5.16

Step (5): Conclusion: Z_{cal}=5.16 Z_{tab} =1.96

If Z_{cal}> Z_{tab} then we reject our H_0.

∴ The samples have not been drawn from same population of S.D 2.5 inches.

Test of significance of single proportion: Suppose a large random sample of size n has a sample proportion p of members possessing a certain attribute. To test the hypothesis that the proportion P in the population has a specified value p_0.

Step (1): Null hypothesis:H_0:p=p_0

Step (2): Alternative hypothesis: H_1:p≠ p_0 / H_1:p<p_0 / H_1:p>p_0

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: Z= $\frac{p-P}{\sqrt{\frac{PQ}{n}}}$

Where p=sample proportion=x/n

P=population proportion, Q=1-P

N=sample size

Step (5): Conclusion: Z_{cal} is compare with Z_{tab} value.

If Z_{cal}<Z_{tab} accept H_0. Otherwise reject H_0.

Problem:

In a sample of 1000 people in Karnataka 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% LOS?

Given n=1000

P=sample proportion of rice eaters=540/1000=0.54

P=population proportion of rice eaters=1/2=0.5 Q=1-P=0.5

Step (1): Null hypothesis:H_0: both rice and wheat are equally popular in the state.i.,e P=0.5

Step (2): Alternative hypothesis:

H_1:p≠ 0.5

Step (3): Level of significance: 1% =2.58

Step (4): Test statistic: Z= $\frac{p-P}{\sqrt{\frac{PQ}{n}}}$ =2.532

Step (5): Conclusion: Z_{cal} = 2.532, Z_{tab} =2.58.

If Z_{cal}<Z_{tab} at 1% LOS then we accept H_0.

Test of equality of two proportions:

Let p_1 and p_2 be the sample proportions in two large random samples of sizes n_1 and n_2 drawn from two populations having proportions p_1 and p_2.

To test whether the two samples have been drawn from the same population

Step (1): Null hypothesis: H_0:P_1= P_2

Step (2): Alternative hypothesis:

$$H_1 : P_1 \neq P_2 \,/\, H_1 : P_1 \leq P_2 \,/\, H_1 : P_1 \geq P_2$$

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: **(a)** when the population proportion P_1 and P_2 are known.

The test statistic is Z=$\frac{p_1 - p_2}{S.E(p_1 - p_2)} \sim N(0,1)$

Where S.E (p_1-p_2) = $\sqrt{\frac{P_1 Q_1}{n_1}} + \sqrt{\frac{P_2 Q_2}{n_2}}$ Q_1=1-P_1 Q_2=1-P_2

(b) When the population proportion P_1 and P_2 are unknown.

In this case we have two methods to estimate P_1 and P_2 .

(i)Method of substitution:

In this method sample proportion p_1 and p_2 are substituted for P_1 and P_2.

$\therefore$ S.E (p_1 - p_2)= $\sqrt{\frac{p_1 q_1 + p_2 q_2}{n_1 + n_2}}$

$\therefore$ Test statistic is Z= $\frac{p_1 - p_2}{S.E\ (p_1 - p_2)}$

(ii) Method of pooling:

In this method, the estimate value for the two population proportions is obtained by pooling the two sample proportions p_1 and p_2 into a single proportion p by the formula is given below.

Sample proportion of two samples or estimated values is given by

P= $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$\therefore$ Test statistic is Z= $\frac{p_1 - p_2}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}}$

Step (5): Conclusion: Z_{cal} is compare with Z_{tab} value.

If Z_{cal}<Z_{tab} accept H_0. Otherwise reject H_0.

Problem:

In two large populations, there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

Given n_1=1200 n_2 =900

p_1= proportion of fair haired people in first population=30/100=0.3

p_2= proportion of fair haired people in first population=25/100=0.25

Step (1): Null hypothesis:H_0: The two sample proportions are equal p_1=p_2

Step (2): Alternative hypothesis:

$$H_1: p_1 \neq p_2$$

Step (3): Level of significance: 5% = 1.96

Step (4): Test statistic: Z= $\frac{p_1 - p_2}{S.E(p_1 - p_2)}$= 2.56

Step (5): Conclusion: Z_{cal} = 2.56, Z_{tab} = 1.96

If Z_{cal}> Z_{tab} at 5% LOS then we reject H_0.

Problem:

(1) Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women in favour of the proposal. Test the hypothesis that proportion of men and women in favour of the proposal at 5% LOS?

Given n_1=400 n_2 =600

p_1= population of men =200/400=0.5

p_2= population of women=250/600=0.541

Step (1): Null hypothesis:H_0: There is no significance difference between the option of men and women $H_0: p_1$=p_2=p

Step (2): Alternative hypothesis:

$$H_1: p_1 \neq p_2$$

Step (3): Level of significance: 5% = 1.96

Step (4): Test statistic: Z= P= $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

∴ Test statistic is Z= $\frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ =1.28

Step (5): Conclusion: Z_{cal} = 1.28, Z_{tab} = 1.96

If Z_{cal}< Z_{tab} at 5% LOS then we reject H_0.

Assignment-cum-Tutorial Questions

SECTION-A

1. Critical region is also knows as _____
2. Whether a test is one-sided or two sided depends on _____ hypothesis.
3. A hypothesis is false, but accepted, this is an error of type_____
4. Rejecting H_0 when H_0 is true is_____ error.
5. The hypothesis which is under test for possible rejection is called _____ hypothesis.
6. A hypothesis contrary to null hypothesis is known as _____hypothesis.
7. Area of critical region depends on
   (a) Size of Type-I error (b) size of Type-II error
   (c) Value of the statistic (d) No. of observations
8. Test of hypothesis H_0:μ=1500 against $H_1 < 1500$ leads to
   (a) One sided lower tailed test (b) one sided upper tailed test
   (c) Two tailed test (d) all the above
9. Level of significance is the probability of
   (a) Type-I error (b) Type-II error (c) both I and II (d) None
10. Among 900 people in a state 90 are found to be chapatti eaters. The 99% confidence interval for the true proportion is
    (a) (0.08, 0.12) (b) (0.8, 1.2) (c) (0.07, 0.13) (d) None
11. Testing H_0:μ=1500 against H_1:μ>1500 leads to:
    (a) One-sided lower tailed test (b) one-sided upper tailed test
    (C) two-tailed test (d) all the above
12. Two samples, one from urban and the other from rural adult males of sizes 400 and 600 had S.D's 165 cm and 175 cm respectively. Test of hypothesis of equality of standard deviations in the two populations at 5% level is:
    (a) Accepted (b) rejected (c) no decision about H_0 (d) none of the above

SECTION-B

1. Explain One-tailed and two-tailed tests.
2. Define Type-I and type-II errors?
3. Explain the procedure for Testing of Hypothesis?
4. Define (a) Critical region (b) Level of significance (c) Left one tailed (d) Right one tailed.
5. The mean life of a sample of 1000 electric bulbs produced by a company is found to be 1570hrs with a S.D of 1200hrs. If μ is the mean life time of all the bulbs produced by the company, test the hypothesis μ=1600hrs against the alternative $\mu \neq$ 1600hrs at 5% LOS.

6. In a random sample of 60 workers the average time taken by them to get to work is 33.8 minutes with a S.D of 6.1 minutes. Can we reject the null hypothesis in favour of alternative hypothesis $\mu > 32.6$ at α=1% LOS.
7. A sample of 900 members has a mean of 3.4cms and S.D 2.61cms. Is the sample from a large population of mean 3.25cms and S.D 2.61cms? If the population is normal and its mean is unknown find the 95% fiducial limits of true mean.
8. Given the following information relating to two places A & B. Test whether there is any significant difference between their mean wages.

	A	B
Mean wages(Rs)	47	49
S.D(Rs)	28	40
No. of workers(Rs)	1000	1500

9. The means of two large samples of sizes 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of S.D 2.5 inches?
10. 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?
11. In a random sample of 400 industrial accidents, it was found that 231 were due at least partially to unsafe working conditions. Construct a 99% confidence interval for the corresponding true proportion.
12. A machine produced 20 defective articles in a batch of 400. After overhauling it produced 10 defectives in a batch of 300. Has the machine improved?

UNIT – V: TESTING OF HYPOTHESIS (SMALL SAMPLES)

Objectives:

- Know principles of hypothesis testing in case of small samples.
- Understand single-sample t-test.
- Understand independent-samples t-test
- Understand how to use an *F*-test to judge whether two population variances are equal.
- Carry out the chi-square test and interpret its results

Syllabus:

t-test, F-test and χ^2 test (independence of attributes and goodness of fit)

Learning Outcomes: The students will be able to

- apply a range of statistical tests appropriately.
- conduct single-sample t-test.
- conduct independent-samples t-test.
- conduct paired (dependent) t-test
- apply *F*-test to judge whether two population variances are equal.
- carry out the chi-square test and interpret its result

In the earlier chapter, we discussed certain tests which are valid only for large samples and hence based on the theory of the Normal distribution and literature on Tests of hypothesis like null hypothesis, Alternative hypothesis, Simple and composite hypothesis, Acceptance and rejection (critical) regions, level of significance, one tailed and two tailed tests etc.

In this chapter, apart from all the above said topics, in addition we need to concentrate on 'degrees of freedom'.

Degrees of freedom: It is very clear that in a test of hypothesis, a sample is drawn from the population of which the parameter is under test. The size of the sample varies since it depends either on the experimenter or on the resources available. Moreover, the test statistic involves the estimated value of the parameter which depends on the number of observations. Hence the sample size plays an important role in testing of hypothesis and is taken care of by degrees of freedom.

Definition: The number of independent observations in a set is called degrees of freedom. It is denoted by ν (read as Nu). In general, the number of degrees of freedom is equal to the total number of observations less than the number of independent constraints imposed on the observations. i.e. in a set of n observations, if k is the number of independent constraints then $\nu = n - k$.

Before going to discuss the tests of significance under small samples, we need some knowledge about exact sampling distributions: t- distribution (or Student's t- distribution),

F- distribution and χ^2 - distribution (or Chi-Square distribution).

t- distribution: It is discovered by W.S.Gosset in 1908. The statistician Gosset is better known by the pen name (pseudonym) 'student' and hence t- distribution is called student's t- distribution.

In practice, the standard deviation σ is not known and in such a situation the only alternative left is to use S, the sample estimate of standard deviation σ. Thus, the variate $\frac{\bar{x}-\mu}{S/\sqrt{n}}$ is approximately normal provided n is sufficiently large. If n is not sufficiently large (small) the varite $\frac{\bar{x}-\mu}{S/\sqrt{n}}$ is distributed as t and hence, $t=\frac{\bar{x}-\mu}{S/\sqrt{n}}$ where $S^2=\frac{1}{(n-1)}\sum_i (x_i-\bar{x})^2$.

Properties of t- distribution:

(1) The shape of t- distribution is bell-shaped and is symmetrical about mean.
(2) The curve of t- distribution is asymptotic to the horizontal axis.
(3) It is symmetrical about the line t = 0.
(4) The form of the probability curve varies with degrees of freedom.
(5) It is unimodel with mean = median = mode.
(6) The mean of t- distribution is zero and variance depends upon the parameter ν , is called the degrees of freedom.
(7) The t- distribution with ν degrees of freedom approaches standard normal distribution as $\nu \to \infty$, ν being a parameter.

The t- distribution is extensively used in hypothesis about one mean or single mean, or about equality of two means or difference of means when σ is known.

Some applications of t- distribution are:

(1). To test the significance of the difference between two sample means or to compare two samples.
(2). To test the significance of an observed sample correlation coefficient and sample correlation coefficient.
(3). To test the significance of difference between two sample means or to compare two samples.
Assumptions about t- test: t- test is based on the following five assumptions.
(1). The random sample has been drawn from a population.
(2). All the observations in the sample are independent.
(3). The sample size is not large. (One should note that at least five observations are desirable for applying a t- test.)
(4). The assumed value μ_0 of the population mean is the correct value.
(5). The sample values are correctly taken and recorded.
(6).The population standard deviation σ is unknown

In case the above assumptions do not hold good, the reliability of the test decreases.

Problem: The life expectancy of people in the year 1970 in Brazil is expected to be 50 years. A survey was conducted in 11 regions of Brazil and the data obtained are given below. Do the data confirm the expected view?
Life expectancy (in years): 54.2, 50.4, 44.2, 49.7, 55.4, 57.0, 58.2, 56.6, 61.9, 57.5, 53.4.

Solution:
Null hypothesis, H_0: $\mu = 50$.

Alternative hypothesis, H_1: $\mu \neq 50$ (Two-tailed test).

Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

Where $\bar{x} = \frac{\sum_i x_i}{n}$, $d_i = x_i - y_i$ and $S^2 = \frac{1}{(n-1)}\sum_i (x_i - \bar{x})^2$.

Calculation of $\bar{x}$ and S^2:

$$\bar{x} = \frac{54.2+50.4+44.2+49.7+55.4+57.0+58.2+56.6+61.9+57.5+53.4}{11} = \frac{598.5}{11} = 54.41.$$

and

$$S^2 = \frac{(54.2-54.41)^2 + (50.4-54.41)^2 + (44.2-54.41)^2 + ... + (53.4-54.41)^2}{10} = \frac{236.07}{10} = 23.607$$

$$\Rightarrow S = 4.853.$$

Therefore, the value of test statistic is

$$t = \frac{54.41-50}{4.859/\sqrt{11}} = 3.01.$$

t- critical value or table value at 5% level of significance and 10 degrees of freedom is 2.228 (from t- tables).

Since t- calculated value is greater than t- critical value, we reject the null hypothesis, H_0. i.e. we accept H_1. It means that the life expectance more than 50 years.

Problem: Mean life time of computers manufactured by a company is 1120 hours. (a) Test the hypothesis that mean lifetime of computers has not changed if a sample of 8 computers has a mean lifetime of 1070 hours with a standard deviation of 125 hours. (b) Is there decrease in mean lifetime? Use (i) 0.05 and (ii) 0.01 level of significance.

Solution:
(a) Null hypothesis, H_0: $\mu = 1120$.

Alternative hypothesis, H_1: $\mu \neq 1120$ (Two-tailed test).

Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{(n-1)}$$

Where $\bar{x} = \frac{\sum_i x_i}{n}$ and $s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$, a sample variance.

Calculation: We are given n = 8, $\bar{x} = 1070,\ s = 125$

Therefore, the value of test statistic is

$$t = \frac{1070 - 1120}{125/\sqrt{7}} = -1.05$$

$or\ |t| = 1.05.$

(i) t- critical value or table value at 5% level of significance with 7 degrees of freedom is 2.365

Since t- calculated value is less than t- critical value, we accept the null hypothesis, H_0.
i.e.the sample has been come from the population whose mean life-time of computers is 1120 hours.

(ii) t- critical value or table value at 1% level of significance with 7 degrees of freedom is 3.499 and we accept null hypothesis H_0.

(b) Null hypothesis, H_0: $\mu = 1120$.

Alternative hypothesis, H_1: $\mu < 1120$ (left-tailed test).

Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{(n-1)}$$

Where $\bar{x} = \frac{\sum_i x_i}{n}$ and $s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$, a sample variance.

Calculation: We are given n = 8, $\bar{x} = 1070,\ s = 125$

Therefore, the value of test statistic is

$$t = \frac{1070 - 1120}{125/\sqrt{7}} = -1.05$$

$or\ t = -1.05$

(i) t- critical value at 5% level of significance with 7 degrees of freedom for left tailed test is -1.895. Since t- calculated value is greater than t- critical value, we accept the null hypothesis, H_0.
i.e .the sample has been come from the population whose mean life-time of computers is 1120 hours.

(ii) t- critical value at 1% level of significance with 7 degrees of freedom is -2.998.

Since t- calculated value is greater than t- critical value, we accept the null hypothesis. i.e. it indicates no decrease in mean lifetime at either of the level of significances.
(These are the applications of t- test for single mean.)

Confidence limits: For example, 95% confidence limits for the population mean μ are $\bar{x} \pm t_\alpha \frac{S}{\sqrt{n}}$ or $\bar{x} \pm t_\alpha \frac{s}{\sqrt{n-1}}$ where $\alpha = 0.025$ for two-tailed test and $\alpha = 5\%$ for one-tailed test, as mentioned in the above example. i.e. For two-tailed test at α los, the value of is taken for $\alpha/2$ from statistical tables of t.

t-test for difference of means:
Assumptions:(i) parent populations , from which the samples have been drawn are normally distributed

(ii) The population variances are equal and unknown
(iii) The two samples are random and independent of each other

Suppose we want to test if two independent samples xi (i=1,2..n1) and yj(j=1,2..n2) of sizes n1 and n2 have been drawn from two normal populations with μ_x and μ_y respectively.

Under the null hypothesis H_0, $\mu_x = \mu_y$, H_1: $\mu_x \neq \mu_y$,

then test statistic

$$t = \frac{\bar{x} - \bar{y}}{S/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

where $\bar{x} = \frac{\sum_i x_i}{n}, \bar{y} = \frac{\sum_i y_i}{n}$ and $S^2 = \frac{1}{(n_1+n_2-2)}\left[\sum_i (x_i - \bar{x})^2 + \sum_i (y_i - \bar{y})^2\right]$, a combined sample mean square.

Problem: A group of 5 patients with medicine A weigh 42, 39, 48, 60, and 41 kilograms. Another group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69 and 62 kilograms. Do you agree with the claim that medicine B increase the weigh significantly?

Solution: Null hypothesis, H_0: There is no significant difference between the medicines A and B with reference to their effect on increase in weight. i.e. $\mu_x = \mu_y$.

Alternative hypothesis, H_1: $\mu_x < \mu_y$ (left-tailed test).

Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

where $\bar{x} = \frac{\sum_i x_i}{n}$ and $S^2 = \frac{1}{(n_1+n_2-2)}\left[\sum_i (x_i - \bar{x})^2 + \sum_i (y_i - \bar{y})^2\right]$, a combined sample mean square.

Calculation of $\bar{x}$ and S^2:

We are given $n_1 = 5, \quad n_2 = 7$.

From the given data,

$$\bar{x} = \frac{42+39+48+60+41}{5} = \frac{230}{5} = 46,$$

$$\bar{y} = \frac{38+42+56+64+68+69+62}{7} = \frac{399}{7} = 57$$

and

$$\sum_{i=1}^{5} (x_i - 46)^2 = (42-46)^2 + (39-46)^2 + ... + (41-46)^2 = 290$$

$$\sum_{i=1}^{7} (y_i - 57)^2 = (38-57)^2 + (42-57)^2 + ... + (62-57)^2 = 926$$

Hence,

$$S^2 = \frac{290+926}{5+7-2} = 121.6$$

$$\Rightarrow S = 11.03.$$

Therefore, the value of test statistic is

$$t = \frac{46-57}{11.03\sqrt{\frac{1}{5} + \frac{1}{7}}} = -\frac{11}{6.46} = -1.7$$

$$\therefore t = -1.7$$

t- Critical value at 5% los with 10 degrees of freedom for right tailed test is -1.812.

Clearly, t- calculated value is greater than t- critical value at 5% los, we accept the null hypothesis. i.e. the medicines A and B do not differ significantly with reference to their effect on increase in weight.

(This is an application of Test for difference of means.)

Problem: Memory capacity of 10 students was tested before and after training. State whether the training was effective or not from the following scores:

Before training:	12	14	11	8	7	10	3	0	5	6
After training:	15	16	10	7	5	12	10	2	3	8

Solution:

Null hypothesis, H_0: There is no significant effect of the training on memory capacity of the students. i.e. $\mu_1 = \mu_2$.

Alternative hypothesis, H_1: Memory capacity of the students has been increased or improved after training. i.e. $\mu_1 < \mu_2$ (left-tailed test).

Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$$

Where $\bar{d} = \frac{\sum_i d_i}{n}$, $d_i = x_i - y_i$ and $S^2 = \frac{1}{(n-1)}\sum_i (d_i - \bar{d})^2$.

Calculation of $\bar{d}$ and S^2:

Let memory capacity before training and after training be x and y respectively.

$$\bar{d} = \frac{(12-15)+(14-16)+(11-10)+(8-7)+(7-5)+(10-12)+(3-10)+(0-2)+(5-3)+(6-8)}{10}$$

$$= -\frac{12}{10} = -1.2$$

and

$$S^2 = \frac{(-3-(-1.2))^2 + (-2-(-1.2))^2 + (1-(-1.2))^2 + ... + (-2-(-1.2))^2}{9} = 7.73$$

$$\Rightarrow S = 2.78$$

Therefore, the value of test statistic is

$$t = \frac{-1.2}{2.78/\sqrt{10}} = -1.365.$$

t- Critical value at 5% los with 9 degrees of freedom for left-tailed test is -1.833.

Since t- calculated value is greater than t- critical value at 5% los with 9 degrees of freedom for left tailed test, we accept the null hypothesis, H_0 i.e. we conclude that there is no change in memory capacity after the training programme (or) there is no use of training programme.

F-distribution :- "The ratio of two sample variances is distributed of F." F-distribution was worked out by G.W. Snedecor and as a mark of respect for Sir R.A.Fisher (Father of modern statistics). Who was defined a statistics Z which is based upon the ratio of two –sample variances initially and hence it is denoted by F. (The first letter of Fisher).

Let S_1^2 be the sample variance of an independent sample of size n_1 drawn from a normal population N $\left(\mu_1, \sigma_1^2\right)$. Similarly , let S_2^2 be the sample variance in an independent sample of size n2 drawn from another normal population N $\left(\mu_2, \sigma_2^2\right)$.Thus S_1^2 and S_2^2 are the variances of two random samples of sizes n_1 and n_2 respectively drawn from two normal populations. In order

to determine whether the two samples came from two populations having equal variances of the two independent random samples defined by

F= $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}=\frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2}$

Which is an F-distribution with v1=n1-1 and v2=n2-1 degrees of freedom.

Properties of F-distribution:

(1) F-distribution curve extends on abscissa from 0 to ∞.

(2) It is an unimodel curve and its mode lies on the point

F= $\frac{k_2(k_1-2)}{k_1(k_2+2)}$ or $\frac{v_2(v_1-2)}{v_1(v_2+2)}$ which is always less than unity

(3) F-distribution curve is a positive skew curve. Generally, the F-distribution curve is highly positive skewed where v2 is small

(4) The mean and variance are defined when v2 ≥ 3 and v2≥5 respectively.

(5) There exists a very useful relation for interchange of degrees of freedom v1 and v2 i.e

$F_{1-\alpha}(v_1,v_2)=\frac{1}{F_\alpha(v_2,v_1)}$

(6) The moment generating function of F-distribution does not exist.

F-test is used to

(1)Test the hypothesis about the equality of two population variances.

(2) test the hypothesis about the equality of two or more population means.

F-test for equality of two population variances: Suppose we want to test whether two independent samples xi (i=1,2..n1) and yj(j=1,2..n2) of sizes n1 and n2 have been drawn from two normal populations with the same variance or not then

Null hypotheses, $H_0:\sigma_x^2=\sigma_y^2$, $H_1:\sigma_1^2\neq\sigma_2^2$.

Under the null hypothesis, H_0, the test statistics is:

$$F=\frac{s_x^2}{s_y^2}\sim F_{(v_1,v_2)} \quad \text{(OR)} \quad F=\frac{s_y^2}{s_x^2}\sim F_{(v_2,v_1)}$$

When $s_x^2>s_y^2$ OR $s_y^2>s_x^2$ respectively

Where $s_x^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\bar{x})^2$ with $\bar{x}=\frac{\sum_{i-1}^{n_1}x_i}{n_1}$

And $s_y^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(y_i-\bar{y})^2$ with $\bar{y}=\frac{\sum_{i-1}^{n}y_i}{n_2}$

Besides t-test , we can also apply a F-test for testing equality of two population means.

F-distribution is a very popular and useful distribution because of its utility in testing of hypothesis about the equality of several population means, two population variances and several regression coefficients in multiple regression coefficient etc.,
As a matter of fact , F-test is the backbone of analysis of variance(ANOVA)

Note: (1) F determines whether the ratio of two sample variances s1 and s2 is too small or too large.

(2) when F is close to 1, the two sample variances s1 and s2 are likely same

(3) F-distribution also known as variance ratio distribution

(4) Dividing S_1^2 and S_2^2 by their corresponding population variances standardizes the sample variance, and hence on theaverage both numerator and denominator approach. Therefore, its customer, to take the large sample variance as the numerator.

(5) F-distribution depends not only on the two parameters, $V_1$and V_2 but also on the order in which they are slated.

Problem: Life expectancy in 9 regions of Brazil in 1900 and in 11 regions of Brazil in 1970 was as given in the table below:

(Source: The review of income and wealth, June 1983)

Regions / Life Expectancy	1	2	3	4	5	6	7	8	9	10	11
1900	42.7	43.7	34.0	39.2	46.1	48.7	49.4	45.9	55.3	-	-
1970	54.2	50.4	44.2	49.7	55.4	57.0	58.2	56.6	61.9	57.5	53.4

It is desired to confirm, whether the variation in life expectancy in various reigns in 1900 and in 1970 in same or not.

Solution: Let the populations in 1900 and in 1970 be considered as $N(\mu_1, \sigma_1^2)$ and $N(\mu_1, \sigma_1^2)$ respectively.

Null hypotheses, H_0 : The variation of life expectancy in various regions in 1900 and in 1970 is same. $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Under the null hypothesis, H_0, the test statistics is :

$$F = \frac{s_1^2}{s_2^2} \sim F_{(v_1, v_2)} \quad \text{(OR)} \quad F = \frac{s_2^2}{s_1^2} \sim F_{(v_2, v_1)}$$

When $s_1^2 > s_2^2$ OR $s_2^2 > s_1^2$ respectively

Where $s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_i - \bar{x}\right)^2$ with $\bar{x} = \frac{\sum_{i-1}^{n_1} x_i}{n_1}$

And $s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(y_i - \bar{y}\right)^2$ with $\bar{y} = \frac{\sum_{i-1}^{n} y_i}{n_2}$

Calculation: $\bar{x} = \frac{405}{9} = 45,\ \bar{y} = \frac{598.5}{11} = 5441$(approximate)

$\sum_{i=1}^{9}\left(x_i - 45\right)^2$ = 5.29+1.69+121+33.64+1.21+13.69+0.9+106.9

= 288.51+19.36=302.87

$\Rightarrow s_1^2 = \frac{302.87}{8} = 37.85$

$\sum_{i=1}^{11}\left(y_i - 54.41\right)^2$ = 0.04+16+104.04+22.09+1+6.76+14.44+4.84+56.25+9.61+1 =236.07

$\Rightarrow s_2^2 = \frac{236.07}{10} = 23.607$

Since $s_1^2 > s_2^2$, the value of test statistics is:

$$F = \frac{37.85}{23.607} = 1.603$$

The table value of F at 5% los with (8, 10) degrees of freedom for two tailed test is 3.85 (From F-tables).

Since F-Calculated value is less than f-tabulated value, we accept H_0. i.e. The sample data confirms the equality of variances in 1900 and 1970 in various regions of brazil or $\sigma_1^2 = \sigma_2^2$.

Practice.

Problem: The house-hold net expenditure on health care in south and north India, in two samples of households, expressed as percentage of total income is shown the following table:

South:	15.0	8.0	3.8	6.4	27.4	19.0	35.3	13.6	
North:	18.8	23.1	10.3	8.0	18.0	10.2	15.2	190.0	20.2

Test the equality of variances of households net expenditure on health care in south and north India.

Problem: The time taken by workers in performing a job by method I and Method II is given below.

Method I	20	16	26	27	23	22	-
Method II	27	33	42	35	32	34	38

Do the data show that the variances of time distribution of population from which these samples are drawn do not differ significantly?

Solution:

Null hypothesis, H_0: There is no significant difference between the variances of time distribution of populations. i.e. $\sigma_1^2 = \sigma_2^2$.

Alternative hypothesis, H_1: $\sigma_1^2 \neq \sigma_2^2$ (Two-tailed test)

Level of significance : Choose $\alpha = 5\% = 0.05$

Under the null hypothesis, H_0, the test statistics is :

$$F = \frac{s_1^2}{s_2^2} \sim F_{(v_1, v_2)} \quad \text{(OR)} \quad F = \frac{s_2^2}{s_1^2} \sim F_{(v_2, v_1)}$$

When $s_1^2 > s_2^2$ OR $s_2^2 > s_1^2$ respectively

Calculation: we are given $n_1 = 6$, $n_2 = 7$

$$\bar{x} = \frac{134}{6} = 22.3, \ \bar{y} = \frac{241}{7} = 34.4$$

$$\sum_{i=1}^{6} (x_i - 22.3)^2 = 81.34, \ \sum_{i=1}^{7} (y_i - 34.4)^2 = 133.72$$

$$\therefore s_1^2 = \frac{81.34}{5} = 16.26 \ and \ s_2^2 = \frac{133.72}{6} = 22.29$$

The value of test statistics is

$$F = \frac{22.29}{16.26} = 1.3699 \cong 1.37$$

F-critical value at 5% los with (5, 6) degrees of freedom for two tailed test is 4.39 (From F-tables)

Since F-Calculated value is less than F-tabulated value t 5% los, we accept H_0. i.e. there is no significant deference between the variances of the time distribution by the workers.

Problem: The nicotine contents in milligrams in two samples of tobacco were found to be as follows:

Sample A	24	27	26	21	25	-
Sample B	27	30	28	31	22	36

Can it be said that the two samples have come from the same normal population?

Hint: When testing the significance of the difference of the means of two samples, we assumed that the two samples came from the same population or from populations with same variances. If the variances of the population are not equal, a significant difference in the

means may arise. Hence, to test the two samples have come from the same population or not, we need to apply both t-test and F-test. But here we note that first apply F-test, as usual manner.

χ^2- distribution: Chi-square distribution was first discovered by Helmert in 1876 and later independently given Karl Pearson in 1900. The χ^2-distribution was discovered mainly as a measure of goodness of fit in case of frequency, distribution, i.e. whether the observed frequencies follow a postulated distribution or not.

If $X_1, X_{2,\ldots\ldots\ldots}, X_n$ are n independent normal variates with mean zero and variance unity, the sum of squares of these variegates is distributed as chi-square with n degrees of –freedom.

Note:

F – Distribution as a special case of Beta dist.

χ^2 – distribution as a special case of Gamma dist.

χ^2 distribution used as non-parameter test whereas t and F distribution are parameter test.

Properties of Ch-Square distribution:

1. The χ^2 – distribution curve lies in the first quadrant since the range of X^2 is from 0 to ∞.
2. The χ^2 – distribution curve is not symmetrical and is highly positive skewed.
3. χ^2 – distribution has only one parameter v, the degrees of freedoms.
4. χ^2 –distribution curve is an unimodel curve and its mode is at the point χ^2 = (v-1).
5. The mean and variance of X^2 –distribution are v and 2v respectively.
6. The moment generating function for chi-square distribution is $M_{x^2}(t) = (1-2t)^{-v/2}$ where v = n-1.
7. Additive property holds good for any number of independent χ^2 – variates.

Application of χ^2 – test: The chi-square test is applicable

1. To test the hypothesis of the variance of population.
2. To test the goodness of fit of the theoretical distribution to observed frequency distribution, in one way classification having k-categories.
3. To test the independence of attributes, when the frequencies are presented in a two way classification (Called the contingency table) etc.,

Conditions for validity of χ^2 – test:

1. Sample size n should be large i.e. $n \geq 50$
2. If individual frequencies o (i=1,2……,n) are small say less than 10 then combine neighboring frequencies (pooling) so that combined frequency O_i is greater than 10.
3. The number of classes' k should be independent.
4. The constraints on the cell frequencies, if any are linear.
5. The constraints on the cell frequencies, if any, are linear.

Problem: The following figures show the distribution of digits in numbers chosen at random from a telephone directory.

Solution:

Null hypothesis, H_0: The digits occur equally frequently in the directory.

Alternative hypothesis, H_1: The digits do not occur equally frequently under the null hypothesis, H_0 the test statistics is,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(n-1)}$$

Where Oi is the observed frequency, Ei is expected frequency

Ei is caliculated as $\sum$oi /n

then expected frequency for each observed frequency Ei=10000/10 =1000

Caliculated χ^2 =58.542 , χ^2 critical value at 5% LOS with 9 d.f is 16.919.

Since Caliculated $\chi^2 > \chi^2$ critical value, reject H_0

The digits do not occur equally frequently in the directory

Problem :A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in second class and 20 got a first class. Do these figures commensurate with the general examination result which is in the ratio 4 : 3 : 2 : 1 for the various categories respectively.

Solution: Null hypothesis, H_0:The observed results commensurate with the general examination results

Alternative hypothesis, $H_{1:}$ The observed results do not commensurate with the general examination results

under the null hypothesis, H_0 the test statistics is,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(n-1)}$$

Where E_i = (no. of share pertaining to Oi/total no. of shares) N

Total no. of shares=10, total frequency N=1000

No. of students who failed, O_1=220

No. of students who secured third class, O_2=170

No. of students who secured second class, O_3=90

No. of students who secured first class, O_4=20

Then E_1= (4/10)500=200, E_2= (3/10)500=150, E_3= (2/10)500=100, E_4= (1/10)500=50

Such that $E_1+E_2+E_3+E_4$=500

Calculated χ^2 =23.667, χ^2 critical value at 5% LOS with 4-1=3 d.f is 7.81

Since Calculated $\chi^2 > \chi^2$ critical value, reject H_0

i.e. The observed results do not commensurate with the general examination results

Problem: Fit a Poisson distribution to the following data and test its goodness of fit at 0.05 L.O.S

X	0	1	2	3	4
F	419	352	154	56	19

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

$= 0,\ otherwise$

Frequency function f(x)=NP(x)

First calculate mean = $\frac{\sum f_i x_i}{\sum f_i} = \lambda$ =0.904

Given frequencies are observed frequencies O_i

Then the expected frequencies E_i's are calculated as

f(x=0)=N x P(x=0)=404.9~405

f(x=1)=N x P(x=1)=366.0296~367

f(x=2)=N x P(x=2)=165.4453~166

f(x=3)=N x P(x=3)=49.8542~50

f(x=4)=N x P(x=4)=11.2670~12

X	0	1	2	3	4
Oi	419	352	154	56	19
Ei	405	367	166	50	12

Null hypothesis, H_0:Poisson distribution is suitable for the given data

Alternative hypothesis, H_1: Poisson distribution is not suitable for the given data

under the null hypothesis, H_0 the test statistics is,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(n-1)}$$

Calculated χ^2 =6.7676, χ^2 critical value at 5% LOS with 5-1-1=3 d.f is 7.82(1 d.f is being lost due to linear constraint,1 d.f is being lost due to estimating the parameter)

Since Calculated $\chi^2 <$ χ^2 critical value, accept H_0.

Problem: Given the following contingency table for hair colour and eye colour. Find the value of χ^2 ? Can we expect good association between hair colour and eye colour?

	Hair colour				
Eye colour		Fair	Brown	Black	Total
	Blue	15	5	20	40
	Gray	20	10	20	50
	Brown	25	15	20	60
	Total	60	30	60	150

Null hypothesis, H_0:The two attributes hair colour and eye colour are independent
Alternative hypothesis, $H_{1:}$ hair colour and eye colour are not independent
under the null hypothesis, H_0 the test statistics is,

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(m-1)\times(n-1)}$$

Where O_{ij} is observed frequency (given)
E_{ij} is expected frequency and calculated as E_{ij} =(i^{th} row total x j^{th} column total) / grand total (N)
E_{11}=(40 x 60) /150=16 similarly E_{12}=8 , E_{13}=16, E_{21}=20, E_{22}=10 , E_{23}=20, E_{31}=24,E_{32}=12, E_{33}=24
Calculated χ^2 =3.6458, χ^2 critical value at 5% LOS with (3-1) (3-1)=4 d.f is 9.488
Since Calculated $\chi^2 < \chi^2$ critical value, accept H_0

i.e., Hair colour and eye colour are independent

Assignment-cum-Tutorial Questions

SECTION:A

1. When d.f. for χ^2 are 100 or more, Chi-square is approximated to _____[ ]
   (a) t-distribution (b) F-distribution (c) Z-distribution (d) none of the above
2. Given the following 8 sample values -4,-3,-3,0,3,3,4,4, the value of student's t-test $H_0{:}\mu = 0$ is_____[ ]
   (a) 2.73 (b) 0.97 (c) 3.30 (d) 0.41

3. If all frequencies of classes are same, the value of χ^2 is_____[ ]
   (a) 1 (b) ∞ (c) 0 (d) none of the above

4. Range of statistic –t is _____[ ]
   (a) -1to 1 (b) -∞ to ∞ (c) 0 to ∞ (d) 0 to 1
5. Range of variance of ratio F is :
   (a) -1to 1 (b) -∞ to ∞ (c) 0 to ∞ (d) 0 to 1
6. In a contingency table, the expected frequencies are computed under _____[ ]
   (a) H_0 (b) H_1 (c) both (a) and (b) (d) no consideration of the hypothesis

7. The shape of t—distribution is similar to that of_____[ ]

(a) Chi-square distribution (b) F-distribution (c) Normal distribution (d) none

8. Which test is used to test the equality of population variances _____[ ]

(a) Chi-square test (b) t-test (c) F-test (d) z-test

9. If two independent random samples of sizes n1=13 and n_2= 7 , are taken from a normal population .The variances of the first sample will be at least four times as that of a second sample then F is _____[ ]

   (a) ¼ (b) 4 (c) 16 (d) none

10. chi-square distribution curve varies from_____[ ]

   (a) - ∞ to ∞ (b) - ∞ to 0 (c) 0 to ∞ (d) none

11. To test the goodness of fit _____test is used [ ]

   (a) z-test (b) F-test (c) χ^2-test (d) t-test

12. Chi-square Coefficient of contingency is calculated when_____[ ]

(a) The attributes are independent (b) the attributes are associated

(c) both (a) and (b) (d) neither (a) nor (b)

13.When the value of coefficient of contingency χ^2 =0, it shows_____[ ]

(a) Complete dissociation amongst attributes
(b) Complete association amongst attributes
(c) Both (a) and (b)
(d) Neither (a) nor (b)

SECTION:B

1. The following are the average weekly losses of worker hours due to accidents in 10 Industrial plants before and after a certain safety programme was put into operation

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

Test whether the safety programme is effective in reducing the number of accidents at the level of significance of 0.05?

2. A machinist is making engine parts with axle diameters of 0.700 inch .A random sample of 10 parts shows a mean diameter of 0.742 inch with a standard deviation of 0.04inch.Compute the statistic you would use to test whether the work is meeting the specification at 0.05L.O.S.

3. A random sample of 6 steel beams has a mean compressive strength of 58,392 p.s.i (pounds per square inch) with a standard deviation of 648 p.s.i. Use this information and the level $\alpha = 0.05$ to test whether the true average compressive strength of the steel from which this sample came is 58,000 p.s.i. Assume normality.

4. A random sample of 10 boys had the following I.Q's:70, 120,110,101,88,83,95,98,107 and 100.
   (a) Do these data support the assumption of a population mean I.Q of 100?
   (b) Find a reason range in which most of the mean I.Q values of samples of 10 boys lie.
5. To examine the hypothesis that the husbands are more intelligent than the wives, an investigator took a sample of 10 couples and administered them a test which measures the I.Q. The results are as follows:

Husbands	117	105	97	105	123	109	86	78	103	107
Wife's	106	98	87	104	116	95	90	69	108	85

Test the hypothesis with a reasonable test at the L.O.S 0.05

6. The blood pressure of 5 women before and after intake of a certain drug are given below :

Before	110	120	125	132	125
After	120	118	125	136	121

Test at 0.05 L.O.S whether there is significant change in B.P.

7. The nicotine content in milligrams in two samples of tobacco were found to be as follows

Sample A	24	27	26	21	25	----
Sample B	27	30	28	31	22	36

Can it be said that the two samples have come from the same normal population?

8. Pumpkins were grown under two experimental conditions. Two random samples of 11 and 9 pumpkins show the sample standard deviations of their weights as 0.8 and 0.5 respectively. Assuming that the weight distributions are normal, test the hypothesis that the true variances are equal.

9. In two independent samples of sizes 8 and 10 the sum of squares of deviations of the sample values from the respective sample means were 84.4 and 102.6. Test whether the difference of variances of the population is significant are not.

10.1000 students at college level were graded to their I.Q and the economic conditions of their homes. Use chi-square test to find out whether there is any association between economic conditions at home and I.Q.

Economic conditions	I.Q		
	High	Low	Total
rich	460	140	600
poor	240	160	400
total	700	300	1000

11.From the following data, find whether there is any significant liking in the habit of taking soft drinks among the categories of employees

Soft drinks	Employees		
	Clerks	Teachers	Officers
Pepsi	10	25	65
Thumsup	15	30	65
Fanta	50	60	30

UNIT – VI
CORRELATION AND REGRESSION

Objectives:

- To understand definition and types of correlation and regression
- To know the determination of correlation coefficient, rank correlation coefficient and regression coefficients

Syllabus:

Types of correlation, Determination of correlation coefficient (for ungrouped data), Rank correlation coefficient, Linear Regression and its properties.

Course Outcomes: After completion of the course the student should be able to

- examine correlation between variables and find the relation between them
- fit the regression lines and forecast
- calculate correlation coefficient, rank correlation coefficient and regression coefficients

Learning Material

Correlation

Correlation is a statistical analysis which measures and analyses the degree or extent to which two variables fluctuate with reference to each other.

The correlation expresses the relationship or interdependence of two sets of variables upon each other. One variable may be called the independent and the other variable dependent etc…

Correlation: If the change in one variable affects the change in the other variable then the two variables are said to be correlated and the relationship is called correlation

Types of Correlation:

Correlation is classified into many types.

- Positive and negative
- Simple and multiple
- Partial and total
- Linear and non-linear

Positive and negative correlation:

If two variables deviate in the same direction i.e. an increase or decrease in the value of one variable is accompanied by an increase or decrease of the other variable, then the correlation is called positive or direct correlation.

If two variables deviate in opposite directions so that an increase or decrease in the values of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative or inverse correlation.

Simple and Multiple correlation:

If the study is between two variables then it **is** simple correlation and examples are quantity of money and price level, demand and price etc,. But in multiple correlations we study more than two variables simultaneously; example is the relationship of price, demand, supply of a commodity.

Partial and Total Correlation:

Two variables excluding some other variables is called partial correlation. Example, we study price and demand, eliminating the supply side. In total correlation, all the facts are taken into account.

Linear and non-linear correlation:

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

In a curvilinear or non-linear correlation, the amount of change in one variable does not bear a constant ration of the amount of change in the other variables.

Scatter diagram or scatter gram:

The scatter diagram is pictorial representation by plotting two variables to find out whether there is any relationship between them.

Karl Pearson's correlation coefficient:

Karl Pearson is a British Biometrician and Statistician suggested a mathematical method for measuring the magnitude of linear relationship between two variables. This is known as Pearson's Coefficient of correlation or Product-Moment correlation coefficient. It is denoted by $r_{x,y}$

$r=\frac{Cov(X,Y)}{\sigma_x \sigma_y}$ OR $r=\frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$ OR $r=\frac{\sum xy}{N\sigma_{x\sigma_y}}$ OR $r=\frac{(\sum XY*n)-(\sum X*\sum Y)}{\sqrt{(\sum X^2*n-(\sum X)^2)*(\sum Y^2*n-(\sum Y)^2)}}$

Where n is number of paired observations

Limits of correlation coefficient ($-1 \leq r_{x,y} \leq +1$)

PROBLEM: Calculate coefficient of correlation from the following data.

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

Solution:

We have $r=\frac{(\sum XY*n)-(\sum X*\sum Y)}{\sqrt{(\sum X^2*n-(\sum X)^2)*(\sum Y^2*n-(\sum Y)^2)}}$

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

Here n=7

$\therefore r=\frac{(676*7)-(70*63)}{\sqrt{(728*7-70^2)*(651*7-63^2)}}$ =0.95

Note: When deviations are taken from an assumed mean the coefficient of correlation is

$$r=\frac{\sum XY-\frac{\sum X \sum Y}{n}}{\sqrt{(\sum X^2-\frac{(\sum X)^2}{n})-(\sum Y^2-\frac{(\sum Y)^2}{n})}}$$

Rank correlation coefficient:

The method of finding the coefficient of correlation by ranks. This method is based on ranks and is useful in dealing with qualitative characteristics such as morality, character, intelligence and beauty. Rank correlation is applicable only to the individual observations. The formula for Spearman's rank correlation coefficient is given by

$$\rho = 1-\frac{6\sum d^2}{n(n^2-1)}$$ (For untied ranks)

Where ρ is rank coefficient of Correlation

d^2 is Sum of the squares of the difference of two ranks

n is Number of paired observations

Properties of rank correlation coefficient:

- **T**he value of ρ lies between 1 and -1
- If ρ=1, there is complete agreement in the order if the ranks and the direction of the rank is same.
- If ρ=-1, then there is complete disagreement in the order of the ranks and they are in opposite directions.

PROBLEM: A random sample of 5 college students is selected and their grades in Mathematics and Statistics are found to be

Mathematics	85	60	73	40	90
Statistics	93	75	65	50	80

Calculate Spearman's rank correlation coefficient.

Solution:

X	Y	Ranks in x	Ranks in y	d_i =x-y	D^2
85	93	2	1	1	1
60	75	4	3	1	1
73	65	3	4	-1	1
40	50	5	5	0	0
90	80	1	2	-1	1
					4

Here N=5 $\sum D^2$=4

Spearman's rank correlation coefficient is

$$\rho = 1-\frac{6\sum d^2}{n(n^2-1)} \quad =1-\frac{6*4}{5(5^2-1)} \quad =0.8$$

Equal or Repeated ranks:

If there is more than one item with the same value in the series then the Spearman's formula for calculating the rank correlation coefficient is

$$\rho = 1-6\left\{\frac{\sum d^2+coreection\ \ factor\ \ of\ X\ and\ Y}{n(n2-1)}\right\}$$

Where correction factor(C.F)= $m(m^2-1)/12$

Where m= the number of times the item is repeated

PROBLEM: Obtain the rank correlation coefficient for the following data

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

X	Y	Rank of X(x)	Rank of Y (y)	d=x-y	d^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				0	72

In X-series, 75 occurs 2 times, so rank = $\frac{2+3}{2}$ =2.5

64 occur 3 times, so rank= $\frac{5+6+7}{3} = 6$

To $\sum d^2$ we add $\frac{m(m^2-1)}{12}$ for each value repeated, so for 75 m=2, for 64, m=3.

So far X series,C.F is $\frac{2(4-1)}{12} + \frac{3(9-1)}{12} = \frac{5}{2}$

In Y series, 68 occurs twice, so rank = $\frac{3+4}{2}$ = 3.5

68 occurs twice so m=2

So far Y series, C.F is $\frac{2(4-1)}{12} = \frac{1}{2}$ $\quad\quad$ $\therefore\ \rho = \frac{1-6(\sum d^2+\frac{5}{2}+\frac{1}{2})}{N(N^2-1)}$=0.545

Regression

In regression analysis the nature of actual relationship if it exists, between two (or more variables) is studied by determining the mathematical equation between the variables. It is mainly used to predict or estimate one (the dependent) variable interns of the other (independent) variable(s).

Simple regression:

It establishes the relationship between two variables (one dependent and one independent variable)

Linear regression:

if the relationship between the two variables is linear and is represented by straight line then it is regression line or the line of average relationship or prediction of equation. Regression lines are of two types (i) regression line y on x (ii) regression line x on y

The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called regression.

Uses:

- It is used to estimate the relation between two economic variables like income and expenditure.
- It is highly valuable tool in economic and business.
- It is useful in statistical estimation of demand curves, supply curves, production function, cost function and consumption function etc.

Deviation taken from arithmetic mean X on Y:

This method is simpler to find the values of a and b. We can find out the deviations of X and Y series from their respective means.

Regression equation X on Y is

$$X-\overline{X} = r\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$$

Regression equation Y on X is

$$(Y-\overline{Y}) = r\frac{\sigma_y}{\sigma_x}(X-\overline{X})$$

Where $\overline{X}$ and $\overline{Y}$ be the means of X and Y series

The regression coefficient of X on Y = $r\frac{\sigma_x}{\sigma_y} = \frac{\sum XY}{\sum Y^2} = b_{xy}$

The regression coefficient of Y on X = $r\frac{\sigma_y}{\sigma_x} = \frac{\sum XY}{\sum X^2} = b_{yx}$

PROBLEM:

Find the most likely production corresponding to a rainfall 40 from the following data.

	Rain fall(X)	Production(Y)
Average	30	500kgs
Standard deviation	5	100kgs
Coefficient of correlation	0.8	

We have to calculate the value of Y when X =40

So we have to find the regression equation of Y on X.

Mean of X series, $\overline{X}$ =30; Mean of Y series, $\overline{Y}$ = 500

σ of X series, σ_x= 5 , σ of Y series , σ_y =100

Regression line Y on X

$$(Y-\overline{Y}) = r\frac{\sigma_y}{\sigma_x}(X-\overline{X}) = (Y-500) = 0.8\left(\frac{100}{5}\right)(X-30)$$

When X=40, Y-500=160

Y=660

Hence the expected value of Y is 660kgs.

Deviations taken from the assumed mean:

If the actual mean is fraction this method is used.

In this method we take deviations from the assumed mean instead of A.M

$$X-\overline{X} = r\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$$

We can find out the value of $r\frac{\sigma_x}{\sigma_y}$ by applying the following formula

$$r\frac{\sigma_x}{\sigma_y} = \frac{\sum dx\ dy - \frac{\sum dx * \sum dy}{n}}{\sum dy^2 - \frac{(\sum dy)^2}{n}}$$, dx =X-A; dy =Y-A

Regression equation Y on X is

$$(Y-\overline{Y}) = r\frac{\sigma_y}{\sigma_x}(X-\overline{X})$$

We can find out the value of $r\frac{\sigma_y}{\sigma_x}$ by applying the following formula

$$r\frac{\sigma_y}{\sigma_x} = \frac{\sum dx\ dy - \frac{\sum dx * \sum dy}{n}}{\sum dx^2 - \frac{(\sum dx)^2}{n}}$$

PROBLEM: Price indices of cotton and wool are given below for the 12 months of a year. Obtain the equations of lines of regression between the indices.

X	78	77	85	88	87	82	81	77	76	83	97	93
Y	84	82	82	85	89	90	88	92	83	89	98	99

Calculation of regression equation

X	dx=(X-84)	dx^2	Y	dy=(Y-88)	dy^2	dxdy
78	-6	36	84	-4	16	24
77	-7	49	82	-6	36	42
85	1	1	82	-6	36	-6
88	4	16	85	-3	9	-12
87	3	9	89	1	1	3
82	-2	4	90	2	4	-4
81	-3	9	88	0	0	0
77	-7	49	92	4	16	-28
76	-8	64	83	-5	25	40
83	-1	1	89	1	1	-1
97	13	169	98	10	100	130
93	9	81	99	11	121	99
1004	-4	488	1061	5	365	287

Regression line X on Y:

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

$$b_{xy} = \frac{\sum dx\ dy - \frac{\sum dx * \sum dy}{n}}{\sum dy^2 - \frac{(\sum dy)^2}{n}} = \frac{287 - (\frac{-4*5}{12})}{365 - \frac{5^2}{12}} = 0.795$$

$$X-83.7 = 0.795(Y-88.42)$$

$$X=0.795Y+13.38$$

Regression line Y on X:

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

$$b_{yx} = \frac{\sum dx\ dy - \frac{\sum dx * \sum dy}{n}}{\sum dx^2 - \frac{(\sum dx)^2}{n}} = \frac{287 - (\frac{-4*5}{12})}{488 - \frac{-4^2}{12}} = 0.59$$

Y-88.42 = 0.59(X-83.67)

Y=0.59X+39.05

PROBLEM:

Determine the equation of a straight line which best fits the data.

X	10	12	13	16	17	20	25
Y	10	22	24	27	29	33	37

Let the required straight line is Y=a+bX

The two normal equations are $\sum Y = b\sum X + na$

$\sum XY = b\sum X^2 + a\sum X$

X	X^2	Y	XY
10	100	10	100
12	144	22	264
13	169	24	312
16	256	27	432
17	289	29	493
20	400	33	660
25	625	37	925
113	1938	182	3186

Substituting the values:

113b+7a =182 ---------- (1)

1983b+113a=3186---------- (2)

Then a=0.82, b=1.56

The equation of straight line is Y=0.82 +1.56 X

UNIT-VI
Assignment-cum-Tutorial Questions

SECTION: A

1. The functional relationship of a dependent variable with independent variable is called _____
2. If there are two or more independent variables in a regression equation, it is named as _____regression.
3. The measure of change in dependent variable corresponding to an unit change in independent variable is called _____
4. The range of Pearson's coefficient of correlation is _____
5. If the correlation coefficient is zero, the value of regression coefficient is _____
6. Scatter diagram of the variate values (X,Y) gives the idea about:
   (a) functional relationship (b) regression model
   (c) distribution of errors (d) none of the above
7. Regression coefficient is independent of:
   (a) Origin (b) scale (c) both (a) & (b) (d) neither (a) nor (b)
8. The range of correlation coefficient is_____
   (a) 0 t o∞ (b) -∞ to ∞ (c) 0 to 1 (d) -1 to 1
9. One regression coefficient is positive then the other regression coefficient is_____
   (a) Positive (b) negative (c) equal to zero (d) cannot say
10. When two regression lines coincide then r is
    (a) 0 (b) -1 (c) 1 (d) 0.5
11. Coefficient of correlation is equal to_____
    (a) $b_{xy} * b_{yx}$ (b) $\sqrt{b_{xy} * b_{yx}}$ (c) $\sqrt{b_{xy}}$ (d) $\sqrt{b_{yx}}$
12. Which of the following indicates the strongest relationship?_____
    (a) $\underline{r} = .5$ (b) $\underline{r} = .09$ (c) $\underline{r} = -.6$ (d) $\underline{r}^2 = .2$
13. In calculating r with raw scores, the numerator of r represents_____
    (a) the variance of X (b) the variance of Y
    (c)the variance of X multiplied by the variance of Y (c)the covariance of X and Y
14. Which of the following would not allow you to calculate a correlation?
    (a)a negative relationship between X and Y (b) a positive relationship between X and Y
    (c)a curvilinear relationship between X and Y (d) a linear relationship between X and Y

SECTION: B

1. Find a suitable coefficient of correlation for the following data:

Fertiliser used(tonnes)	15	18	20	24	30	35	40	50
Productivity(tonnes)	85	93	95	105	120	130	150	160

2. Calculate Karl Pearson's correlation coefficient for the following data.

X	38	45	46	38	35	38	46	32	36	38
Y	28	34	38	34	36	26	28	29	25	36

   What inference would you draw from estimate?
3. Determine Karl Pearson's coefficient of correlation from the data which represents father's height (X) and son's height (Y).

X	64	65	66	67	68	69	70
Y	66	67	65	68	70	68	72

Comment on the result. [II-II Supple June2017 CSE]

4. Given n=10,σ_x = 5.4,σ_y= 6.2 and sum of product of deviation from the mean of X and Y is 66 find the correlation coefficient.
5. Find coefficient of correlation between X and Y for the following data.

X	10	12	18	24	23	27
Y	13	18	12	25	30	10

6. Use the formula $r = \frac{\sigma_X^2+\sigma_Y^2-\sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$ to compute the correlation coefficient to the following data

X	62	56	36	66	25	75	82	78
Y	58	44	51	58	60	68	62	84

[II-II Regular May2016 CSE]

7. Ten competitors in a musical test were ranked by the three judges A,B and C in the following order.

Ranks by A	1	6	5	10	3	2	4	9	7	8
Ranks by B	3	5	8	4	7	10	2	1	6	9
Ranks by C	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pair of judges has the nearest approach to common liking in music.

8. Price indices of cotton and wool are given below for the 12 months of a year. Obtain the equations of lines of regression between the indices.

X	78	77	85	88	87	82	81	77	76	83	97	93
Y	84	82	82	85	89	90	88	92	83	89	98	99

9. Compute the two regression equations from the following data

x	1	2	3	4	5
y	2	3	5	4	6

Estimate the value of y when x = 2.5. [II-II Regular April2017 CSE]

10. Calculate the regression equations of Y on X from the data given below, taking deviations from actual means of X and Y.

Price(Rs.)	10	12	13	12	16	15
Amount Demanded	40	38	43	45	37	43

Estimate the likely demand when the price is Rs. 20.

11. The following calculations have been made for prices of 12 stocks (X) in stock exchange, on a certain day along with the volume of the sales in thousands of shares(Y). From these calculations find the regression equation of prices of stocks, on the volume of the sales of shares.
$\sum X$ =580, $\sum Y$ =370, $\sum XY$ =11499, $\sum X^2$=41658, $\sum Y^2$=17206
12. The equations of two regression lines are 7X-16Y+9=0 and 5Y-4X-3=0. Find the coefficient of correlation and the means of X and Y.
13. The equations of two regression lines obtained in a correlation analysis are $8x - 10y + 66 = 0$, $40x - 18y = 214$ Find (i) mean values of x and y (ii) Correlation coefficient between x and y.

[II-II Regular May2016 CSE]

14. If X=4Y +5 and Y=KX + 4 are the lines of regression of X on Y and Y on X respectively, show that $0 < 4k < 1$. If $= \frac{1}{16}$, find the means of the two variables and the coefficient of correlation between them. [II-II Supple Jan2017 CSE]

@@