PROBABILITY & STATISTICS

Course code:AHSB12

II. B.Tech III semester

Regulation: IARE R-18)

BY
Dr. S. Jagdha
Associate Professor
Ms. P. Srilatha, Ms. B. Praveena & Ms. V. Subbalaxmi
Assistant Professors

DEPARTMENT OF MECHANICAL ENGINEERING
INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)
DUNDIGAL, HYDERABAD - 500 043

CO’s	Course outcomes
CO1	Describe the concept of probability, conditional probability, Baye’s theorem and analyze the concepts of discrete, continuous random variables
CO2	Determine the binomial, poisson and normal distribution to find mean, variance
CO3	Understand multiple random variables and enumerate correlation and regression to the given data.
CO4	Explore the concept of sampling distribution and apply testing of hypothesis for sample means and proportions.
CO5	Use t-test for means, F-test for variances and chi-square test for independence to determine whether there is a significant relationship between two categorical variables.

MODULE– I
PROBABILITY AND RANDOM VARIABLES

CLOs	Course Learning Outcome
CLO1	Describe the basic concepts of probability
CLO2	Summarize the concept of conditional probability and estimate the probability of event using Baye’s theorem.
CLO3	Analyze the concepts of discrete and continuous random variables, probability distributions, expectation and variance.
CLO4	Use the concept of random variables in real-world problem like graph theory; machine learning, Natural language processing.

Exhaustive Events:

The total no. of possible outcomes in any trial is called exhaustive event.

E.g.: (1) In tossing of a coin experiment there are two exhaustive events.

(2) In throwing an n-dice experiment, there are 6^n exhaustive events.

Favorable event:

The no of cases favorable to an event in a trial is the no of outcomes which entities the happening of the event.

E.g. (1) In tossing a coin, there is one and only one favorable case

to get either head or tail.

Mutually exclusive Event:

If two or more of them cannot happen simultaneously in the same trial then the event are called mutually exclusive event.

E.g. In throwing a dice experiment, the events 1,2,3,------6 are M.E. events

Equally likely Events:

Outcomes of events are said to be equally likely if there is no reason for one to be preferred over other. E.g. tossing a coin. Chance of getting 1,2,3,4,5,6 is equally likely.

Independent Event:

Several events are said to be independent if the happening or the non-happening of the event is not affected by the concerning of the occurrence of any one of the remaining events.

An event that always happen is called **Certain event,** it is denoted by ‘S’.

An event that never happens is called **Impossible event**, it is denoted by ‘ϕ’.

Eg: In tossing a coin and throwing a die, getting head or tail is independent of getting no’s 1 or 2 or 3 or 4 or 5 or 6.

Probability (Mathematical Definition)

Definition: **I**f a trial results in n-exhaustive mutually exclusive, and equally likely cases and m of them are favorable to the happening of an event E then the probability of an event E is denoted by P(E) and is defined as

$$P(E) = \frac{no\,of\ favourable\,cases\,to\,event}{Total\,no\,of\ exaustive\,cases} = \frac{m}{n}$$

Sample Space:

The set of all possible outcomes of a random experiment is called Sample Space .The elements of this set are called sample points. Sample Space is denoted by S.

Eg. (1) In throwing two dies experiment, Sample S contains 36 Sample points.

S = {(1,1) ,(1,2) ,----------(1,6), --------(6,1),(6,2),--------(6,6)}

Eg. (2) In tossing two coins experiment , S = {HH ,HT,TH,TT}

A sample space is called **discrete** if it contains only finitely or infinitely many points which can be arranged into a simple sequence w_1,w_2,……. .while a sample space containing non denumerable no. of points is called a continuous sample space.

Statistical or Empirical Probability:

If a trial is repeated a no. of times under essential homogenous and identical conditions, then the limiting value of the ratio of the no. of times the event happens to the total no. of trials, as the number of trials become indefinitely large, is called the probability of happening of the event.(It is assumed the limit is finite and unique)

Conditional Probability

The ***conditional probability*** of an event *B* is the probability that the event will occur given the knowledge that an event *A* has already occurred. This probability is written *P(B|A),* notation for the *probability of B given A*. In the case where events *A* and *B* are *independent* (where event *A* has no effect on the probability of event *B*), the conditional probability of event *B* given event *A* is simply the probability of event *B*, that is *P(B)*.

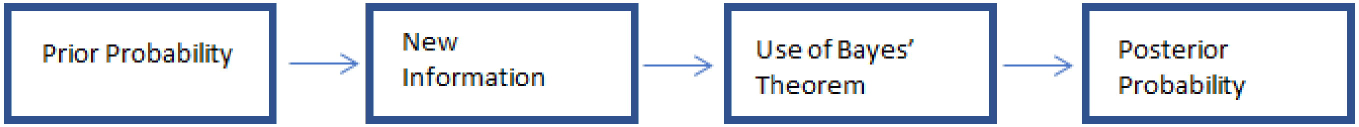
If events *A* and *B* are not independent, then the probability of the *intersection of A and B* (the probability that both events occur) is defined by *P(A and B) = P(A)P(B|A).*

From this definition, the conditional probability *P(B|A)* is easily obtained by dividing by *P(A)*:

$$P(B|A) = \frac{P(A \ and \ B)}{P(A)}$$

Baye’s Theorem

We are quite familiar with probability and its calculation. From one known probability we can go on calculating others. But can we use all the prior information to calculate or to measure the chance of some events happened in past? This is the posterior probability. Bayes theorem calculates the posterior probability of a new event using a prior probability of some events.



Prior Probability

New Information

Use of Bayes' Theorem

Posterior Probability

Baye’s theorem, sometimes, also calculates the probability of some future events.

Theorem

If E_1, E_2, E_3, … , E_n are mutually disjoint events with $P(E_i) \neq 0$, (i = 1, 2, …, n), then for any arbitrary event A which is a subset of the union of events E_i such that $P(A) > 0$, we have

$P(E_i \mid A) = [P(E_i).P(A \mid E_i)] \div [\sum_i P(E_i).P(A \mid E_i)] = [P(E_i).P(A \mid E_i)] \div P(A)$, E_1, E_2, E_3, … , E_n represents the partition of the sample space S

Proof:

A is a subset of the union of E_i, i.e., $A \subset \cup E_i$, we have, $A = A \cap (\cup E_i) = \cup (A \cap E_i)$. Also, $(A \cap E_i)$ subset E_i, $(i = 1, 2, \ldots, n)$ are mutually disjoint events as E_i's are mutually disjoint. So,

$P(A) = P[\cup (A \cap E_i)] = \sum_i P(A \cap E_i) = \sum_i P(E_i). P(A \mid E_i)$.

Also, $P(A \cap E_i) = P(A). P(E_i \mid A)$

or, $P(E_i \mid A) = P(A \cap E_i) \,/\, P(A) = [P(E_i). P(A \mid E_i)] \div [\sum_i P(E_i). P(A \mid E_i)]$.

Problems:

1.Let us consider the situation where a child has three bags of fruits in which Bag 1 has 5 apples and 3 oranges, Bag 2 has 3 apples and 6 oranges and Bag 3 has 2 apples and 3 oranges. One fruit is drawn at random from one of the bags. It was an apple. Let us calculate the probability that the chosen fruit was apple and was drawn from Bag 3.

Here, we can calculate the probability of selecting the bags, $P(E_1) = P(E_2) = P(E_3) = 1/3$. The probability of drawing out of apple from Bag 1, $P(A \mid E_1) = 5/8$, from Bag 2, $P(A|E_2) = 3/9 = 1/3$, from Bag 3, $P(A \mid E_3) = 2/5$.

We have to calculate the probability of drawing a fruit given that we have chosen the Bag 3. The probability of drawing a fruit from Bag 3 given that the chosen fruit is an apple is $P(E_3|A)$. The Baye's formula helps us to calculate the probability, which is

$P(E_3|\ A) = [P(E_3).\ P(A\ |\ E_3)] \div [P(E_1)\times P(A\ |E_1) + P(E_2)\times P(A\ |\ E_2) + P(E_3)\times P(A|E_3)]$

$\Rightarrow P(E_3|\ A)\ 0= 1/3 \times 2/5 \div [(1/3 \times 5/8) + (1/3 \times 3/9) + (1/3 \times 2/5)] = (2/15) \div (163/360) = 48/163.$

2. One box is chosen at random and three balls are drawn from it. They all are of different colors. What is the probability that they come from boxes I, II or III?

Solution: Let A be the event of drawing three balls. E_1, E_2, E_3 represent the events of selecting Box I, II and III respectively.

$P(E_1) = P(E_2) = P(E_3) = 1/3$.

The probability of drawing three balls given that the Box I is selected,

$P(A|E_1) = ({}^1C_1 \times {}^2C_1 \times {}^3C_1) \div {}^6C_3 = 3/10.$

$P(A|E_2) = (1 \times 1 \times 2) \div {}^4C_3 = ½.$

$P(A|E_3) = (5 \times 4 \times 1) \div {}^{10}C_3 = 1/6.$

And $P(E_1)\times P(A|E_1) + P(E_2)\times P(A | E_2) + P(E_3)\times P(A | E_3) = 29/90.$

$P(E_3 \mid A) = 1 - [P(E_1 \mid A) + P(E_2 \mid A)] = 1 - [(9/29) + (15/29)] = 1 - 24/29 = 5/29.$

Random Variables

- A random variable X on a sample space S is a function X : S → R from S onto the set of real numbers R, which assigns a real number X (s) to each sample point ‘s’ of S.
- Random variables (r.v.) bare denoted by the capital letters X,Y,Z,etc..
- Random variable is a single valued function.
- Sum, difference, product of two random variables is also a random variable .Finite linear combination of r.v is also a r.v .Scalar multiple of a random variable is also random variable.

- A random variable, which takes at most a countable number of values, it is called a discrete r.v. In other words, a real valued function defined on a discrete sample space is called discrete r.v.
- A random variable X is said to be continuous if it can take all possible values between certain limits .In other words, a r.v is said to be continuous when it’s different values cannot be put in 1-1 correspondence with a set of positive integers.

- A continuous r.v is a r.v that can be measured to any desired degree of accuracy. Ex : age , height, weight etc..
- Discrete Probability distribution: Each event in a sample has a certain probability of occurrence . A formula representing all these probabilities which a discrete r.v. assumes is known as the discrete probability distribution.

- The probability function or probability mass function (p.m.f) of a discrete random variable X is the function f(x) satisfying the following conditions.

  i) $f(x) \geq 0$

  ii) $\sum_{x} f(x) = 1$

  iii) $P(X = x) = f(x)$

- Cumulative distribution or simply distribution of a discrete r.v. X is F(x) defined by $F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$ for $-\infty < x < \infty$

- For a continuous r.v. X, the function f(x) satisfying the following is known as the probability density function(p.d.f.) or simply density function:

  i) f(x) $\geq$ 0 ,-∞ <x <∞

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

  iii) P(a<X<b)=$\int_{a}^{b} f(x)dx$= Area under f(x) between

  ordinates x=a and x=b

- Cumulative distribution for a continuous r.v. X with p.d.f. f(x), the cumulative distribution F(x) is defined as

F(x)= P(X≤x)= $\int_{-\infty}^{\infty} f(t)dt$ $-\infty<x<\infty$

It follows that F(-∞) = 0 , F(∞)=1, 0≤F(x)≤1

for $-\infty<x<\infty$

f(x)= d/dx(F(x))= $F^1(x)\geq 0$ and P(a < x < b)= F(b)-F(a)

- In case of discrete r.v. the probability at a point i.e., P(x=c) is not zero for some fixed c however in case of continuous random variables the probability at appoint is always zero. I.e., P(x=c) = 0 for all possible values of c.

P(E) = 0 does not imply that the event E is null or impossible event.

- If X and Y are two discrete random variables the joint probability function of X and Y is given by P(X=x,Y=y) = f(x,y) and satisfies

  (i) f(x,y) $\geq$ 0 (ii) $\sum_{x}\sum_{y} f(x,y)$ = 1

The joint probability function for X and Y can be reperesented by a joint probability table.

Table

X Y	y_1	y_2	……	y_n	**Totals**
x_1	$f(x_1,y_1)$	$f(x_1,y_2)$	…….	$f(x_1,y_n)$	$f_1(x_{1)}$ $=P(X=x_1)$

x_2	$F(x_2,y_1)$	$f(x_2,y_2)$	……..	$f(x_2,y_n)$	$f_1(x_2)$ $=P(X=x_2)$
……..	…….	………	………	………	……..
x_m	$f(x_m,y_1)$	$f(x_m,y_2)$	…….	$f(x_m,y_n)$	$f_1(x_m)$ $=P(X=x_m)$
Totals	$f_2(y_1)$ $=P(Y=y_1)$	$f_2(y_2)$ $=P(Y=y_2)$	……..	$f_2(y_n)$ $=P(Y=y_n)$	1

The probability ofX = x_j is obtained by adding all entries in arrow corresponding to X = x_j

Similarly the probability of Y = y_k is obtained by all entries in the column corresponding to Y =y_k

$f_1(x)$ and $f_2(y)$ are called marginal probability functions of X and Y respectively.

The joint distribution function of X and Y is defined by F(x,y)= P(X≤x,Y≤y)= $\sum_{u\leq x}\sum_{v\leq y} f(u,v)$

The probability ofX = x_j is obtained by adding all entries in arrow corresponding to X = x_j

Similarly the probability of Y = y_k is obtained by all entries in the column corresponding to Y =y_k

$f_1(x)$ and $f_2(y)$ are called marginal probability functions of X and Y respectively.

- Two discrete random variables X and Y are independent iff

$$P(X = x, Y = y) = P(X = x)P(Y = y) \;\forall\; x,y \quad \text{(or)}$$

$$f(x,y) = f_1(x)f_2(y) \quad \forall\; x, y$$

- Two continuous random variables X and Y are independent iff

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \;\forall\; x,y$$

(or)

$$f(x,y) = f_1(x)f_2(y) \quad \forall\; x, y$$

If X and Y are two discrete r.v. with joint probability function f(x,y) then

$$P(Y = y|X=x) = \frac{f(x,y)}{f_1(x)} = f(y|x)$$

Similarly, $$P(X = x|Y=y) = \frac{f(x,y)}{f_2(y)} = f(x|y)$$

- Median is the point, which divides the entire distribution into two equal parts. In case of continuous distribution median is the point, which divides the total area into two equal parts. Thus, if M is the median then $\int_{-\infty}^{M} f(x)dx$ = $\int_{M}^{\infty} f(x)dx$=1/2. Thus, solving any one of the equations for M we get the value of median. Median is unique

Mode: Mode is the value for f(x) or P(x_i) at attains its maximum

For continuous r.v. X mode is the solution of $f^1(x) = 0$ and $f^{11}(x) <0$ provided it lies in the given interval. Mode may or may not be unique.

- Variance: Variance characterizes the variability in the distributions with same mean can still have different dispersion of data about their means

Variance of r.v. X denoted by Var(X) and is defined as

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right] = \begin{cases} \sum_{x}(x-\mu)^2 f(x) & \text{for discrete} \\ \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx & \text{for continuous} \end{cases}$$

where μ = E(X)

- If c is any constant then E(cX) = c E(X)
- If X and Y are two r.v.’s then E(X+Y) = E(X)+E(Y)
- IF X,Y are two independent r.v.’s then E(XY) = E(X)E(Y)

- If X_1, X_2,-------,X_n are random variables then $E(c_1X_1 + c_2X_2 +$------$+ c_nX_n) = c_1E(X_1) + c_2E(X_2) +$-----$+ c_nE(X_n)$ for any scalars c_1, c_2,------,c_n If all expectations exists

- If X_1, X_2,-------,X_n are independent r.v's then E $\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i)$ if all expectations exists.

- Var (X) = E (X^2) –[E (X)]^2

- If 'c' is any constant then var (cX) = c^2var(X)

- The quantity $E[(X-a)^2]$ is minimum when a =μ= E(X)

- If X and Y are independent r.v.'s then Var(X $\pm$ Y) = Var(X) $\pm$ Var(Y)

1:A random variable x has the following probability function:

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

Find (i) k (ii) P(x<6) (iii) P(x>6)

Solution:since the total probability is unity, we have $\sum_{x=0}^{n} p(x)=1$

i.e., 0 + k +2k+ 2k+ 3k+ k^2+ $7k^2$+k=1

i.e., $8k^2$+ 9k $-1=0$

$k=1,-1/8$

$P(x<6) = 0 + k + 2k + 2k + 3k$

$= 1+2+2+3 = 8$

iii) $P(x>6) = k^2 + 7k^2 + k$

$= 9$

2. Let X denotes the minimum of the two numbers that appear when a pair of fair dice is thrown once. Determine (i) Discrete probability distribution (ii) Expectation (iii) Variance

Solution:

When two dice are thrown, total number of outcomes is 6x6=36

In this case, sample space S= $\begin{array}{l} \{(1,1)(1,2)(1,3)(1,4)(1,5)(1,6) \\ (2,1)(2,2)(2,3)(2,4)(2,5)(2,6) \\ (3,1)(3,2)(3,3)(3,4)(3,5)(3,6) \\ (4,1)(4,2)(4,3)(4,4)(4,5)(4,6) \\ (5,1)(5,2)(5,3)(5,4)(5,5)(5,6) \\ (6,1)(6,2)(6,3)(6,4)(6,5)(6,6) \end{array}$

If the random variable X assigns the minimum of its number in S, then the sample space S=

$$\left\{\begin{matrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 & 5 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{matrix}\right\}$$

The minimum number could be 1,2,3,4,5,6

For minimum 1, the favorable cases are 11

Therefore, P(x=1)=11/36

P(x=2)=9/36, P(x=3)=7/36, P(x=4)=5/36, P(x=5)=3/36, P(x=6)=1/36

The probability distribution is

X	1	2	3	4	5	6
P(x)	11/36	9/36	7/36	5/36	3/36	1/36

(i)Expectation mean =$\sum p_i x_i$

$$E(x)=1\frac{11}{36}+2\frac{9}{36}+3\frac{7}{36}+4\frac{5}{36}+5\frac{3}{36}+6\frac{1}{36}$$

Or $\mu=\frac{1}{36}\left[11+8+21+20+15+6\right]=\frac{9}{36}=2.5278$

ii) variance = $\sum p_i x^2 i - \mu^2$

$$E(x) = \frac{11}{36}1 + \frac{9}{36}4 + \frac{7}{36}9 + \frac{5}{36}16 + \frac{3}{36}25 + \frac{1}{36}36 - (2.52$$

=1.9713

A continuous random variable has the probability density function

$$f(x) = \begin{cases} kxe^{-\lambda x}, \ for\ x \geq 0, \lambda > 0 \\ 0, otherwise \end{cases}$$

Determine (i) k (ii) Mean (iii) Variance

Solution:

(i) since the total probability is unity, we have $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_{-\infty}^{0} 0dx + \int_{0}^{\infty} kxe^{-\lambda x}dx = 1$$

i.e., $\int_{0}^{\infty} kxe^{-\lambda x}dx = 1$

$$k\left[x\left(\frac{e^{-\lambda x}}{-\lambda}\right) - 1\left(\frac{e^{-\lambda x}}{\lambda^2}\right)\right]_0^{\infty} or\, k = \lambda^2$$

(i) mean of the distribution $\mu = \int_{-\infty}^{\infty} xf(x)dx$

$$\int_{-\infty}^{0} 0dx + \int_{0}^{\infty} kx^2 e^{-\lambda x} dx$$

$$\lambda^2 \left[x^2 \left(\frac{e^{-\lambda x}}{-\lambda} \right) - 2x \left(\frac{e^{-\lambda x}}{\lambda^2} \right) + 2 \left(\frac{e^{-\lambda x}}{\lambda^3} \right) \right]_0^{\infty}$$

$$= \frac{2}{\lambda}$$

Variance of the distribution

$$\sigma 2 = \int_{-\infty}^{\infty} x2\, f(x) dx - \mu 2$$

$$\sigma 2 = \int_{-\infty}^{\infty} x2\, f(x) dx - \frac{4}{\lambda 2}$$

$$= \lambda 2 \left[x3 \left(\frac{e-\lambda x}{-\lambda} \right) - 3x2 \left(\frac{e-\lambda x}{\lambda 2} \right) + 6x \left(\frac{e-\lambda x}{\lambda 3} \right) - 6 \left(\frac{e-\lambda x}{\lambda 4} \right) \right]_0^{\propto}$$

$$= \frac{2}{\lambda 2}$$

Out of 800 families with 5 children each, how many would you expect to have (i)3 boys (ii)5girls (iii)either 2 or 3 boys ? Assume equal probabilities for boys and girls

Solution(i)

P(3boys)=P(r=3)=P(3)=$\frac{1}{2^5}{}^5C_3=\frac{5}{16}$per family

Thus for 800 families the probability of number of families having 3 boys=$\frac{5}{16}\left(800\right)=250$families

P(5 girls)=P(no boys)=P(r=0)= $\frac{1}{2^5}{}^5C_0=\frac{1}{32}$per

Family Thus for 800 families the probability of

number

of families having 5girls=$\frac{1}{32}(800)=25$families

P(either 2 or 3 boys =P(r=2)+P(r=3)=P(2)+P(3)

$\frac{1}{2^5}{}^5C_2+\frac{1}{2^5}{}^5C_3$=5/8 per family

Expected number of families with 2 or 3 boys = $\frac{5}{8}(800)=500$families.

4. In 1000 sets of trials per an event of small probability the frequencies f of the number of x of successes are

X	0	1	2	3	4	5	6	7	Total
P(X)	305	365	210	80	28	9	2	1	1000

of 10 randomly chosen tape recorders. Find (i) P(X=0) (ii) P(X=1) (iii) P(X=2) (iv) P (1<X<4).

MODULE-II

PROBABILITY DISTRIBUTION

MODULE-II

CLOs	Course Learning Outcome
CLO 5	Determine the binomial distribution to find mean and variance.
CLO 6	Understand binomial distribution to the phenomena of real-world problem like sick versus healthy.
CLO 7	Determine the poisson distribution to find mean and variance.
CLO 8	Use poisson distribution in real-world problem to predict soccer scores

MODULE-II

CLOs	Course Learning Outcome
CLO 9	Illustrate the inferential methods relating to the means of normal distributions
CLO 10	Describe the mapping of normal distribution in real-world problem to analyze the stock market.

Bionomial Distribution

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

$$P(X = x) = \begin{cases} P(x) = \binom{n}{x} p^x q^{n-x} & \text{where } x = 0,1,2,3,\ldots n \quad q = 1-p \\ 0 \text{ other wise} & \end{cases}$$

where n, p are known as parameters, n- number of independent trials p- probability of success in each trial, q- probability of failure.

- **Mode of the Binomial distribution**: Mode of B.D. Depending upon the values of (n+1)p

(i) If (n+1)p is not an integer then there exists a unique modal value for binomial distribution and it is ‘m’= integral part of (n+1)p

(ii) If (n+1)p is an integer say m then the distribution is Bi-Modal and the two modal values are m and m-1

Poisson distribution

Poisson distribution:

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

(i) n, the number of trials is infinitely large.

(ii) P, the constant probability of success for each trial is indefinitely small.

(iii) np= λ, is finite where λ is a positive real number.

- A random variable X is said to follow a Poisson distribution if it assumes
- only non-negative values and its p.m.f. is given by

$$P(x,\lambda)= \begin{cases} P(X= x) = \dfrac{e^{-\lambda}\lambda^{x}}{x!} & : \quad x= 0,1,2,3,\ldots\ldots \ \lambda > 0 \\ 0 & \text{Other wise} \end{cases}$$

Here λ is known as the parameter of the distribution.

- We shall use the notation X~ P(λ) to denote that X is a Poisson variate with parameter λ
- Mean and variance of Poisson distribution are equal to λ.
- The coefficient of skewness and kurtosis of the poisson distribution are $\gamma_1 = \sqrt{\beta_1} = 1/\sqrt{\lambda}$ and $\gamma_2 = \beta_2 - 3 = 1/\lambda$. Hence the poisson distribution is always a skewed distribution. Proceeding to limit as λ tends to infinity we get $\beta_1 = 0$ and $\beta_2=3$

Mode of Poisson Distribution: Mode of P.D. Depending upon the value of λ

(i) when λ is not an integer the distribution is uni- modal and integral part of λ is the unique modal value.

(ii) When λ = k is an integer the distribution is bi-modal and the two modals are k-1 and k.

(iii) Sum of independent poisson variates is also poisson variate.

(iii) The difference of two independent poisson variates is not a poisson
(iv) variate.

- **Moment generating function of the P.D.**

  If X~ P(λ) then $M_X(t)$ = $e^{\lambda(e^t-1)}$

- Recurrence formula for the probabilities of P.D. (Fitting of P.D.)

  P(x+1) = $\frac{\lambda}{x+1}p(x)$

- Recurrence relation for the probabilities of B.D. (Fitting of B.D.)

  P(x+1) = $\left\{\frac{n-x}{x+1}.\frac{p}{q}\right\}p(x)$

1. Average number of accidents on any day on a national highway is 1.8. Determine th probability that the number of accidents is (i) at least one (ii) at most one

Solution:

Mean= $\lambda = 1.8$

We have P(X=x)=p(x)$\frac{e^{-\lambda}\lambda^{x}}{x!}$ =$\frac{e^{-1.8}1.8^{x}}{x!}$

i) P (at least one) =P(x≥1)=1-P(x=0)

=1-0.1653

=0.8347

ii) P (at most one) =P (x≤1)

=P(x=0)+P(x=1)

= 0.4628

Exercise problems:

2.If a Poisson distribution is such that $P(X=1)=\frac{3}{2}P(X=3)$ then find

(i) $P(X\geq 1)$

(ii) $P(X\leq 3)$ (iii) $P(2\leq X\leq 5)$.

Normal Distribution

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

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

- A r.v. X with mean μ and variance σ^2 follows the normal distribution is denoted by

  $X \sim N(\mu, \sigma^2)$

- If $X \sim N(\mu, \sigma^2)$ then $Z = \frac{X-\mu}{\sigma}$ is a standard normal variate with $E(Z) = 0$ and $var(Z)=0$ and we write $Z \sim N(0,1)$

- The p.d.f. of standard normal variate Z is given by f(Z) = $\frac{1}{\sqrt{2\pi}} e^{-z^2/2}$,
- -∞ < Z< ∞
- The distribution function F(Z) = P(Z ≤ z) = $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$
- F(-z) = 1 – F(z)
- P(a < z ≤ b) = P(a ≤ z < b)= P(a <z < b)= P(a ≤ z ≤ b)= F(b) – F(a)

- If X~ N(μ, σ^2) then Z = $\frac{X-\mu}{\sigma}$ then P(a $\leq$ X $\leq$ b) = $F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right)$
- N.D. is another limiting form of the B.D. under the following conditions:

  i) n , the number of trials is infinitely large.

  ii)Neither p nor q is very small

i) The maximum probability occurring at the point x= μ and is given by

$[P(x)]_{max} = 1/\sigma\sqrt{2\Pi}$

ii) $\beta_1 = 0$ and $\beta_2 = 3$

$\mu_{2r+1} = 0$ (r = 0,1,2……) and $\mu_{2r} = 1.3.5….(2r-1)\sigma^{2r}$

iii) Since f(x) being the probability can never be negative no portion of the curve lies below x- axis.

i) Linear combination of independent normal variate is also a normal variate.

ii) X- axis is an asymptote to the curve.

iii) The points of inflexion of the curve are given by x = $\mu \pm \sigma$, f(x) = $\frac{1}{\sigma\sqrt{2\pi}}e^{-1/2}$

iv) Q.D. : M.D.: S.D. :: $\frac{2}{3}\sigma$: $\frac{4}{5}\sigma$: σ ::: $\frac{2}{3}$ $\frac{4}{5}$: 1 Or Q.D. : M.D.: S.D. **::10:12:15**

Problems

Problems:

1. The mean weight of 800 male students at a certain college is 140kg and the standard deviation is 10kg assuming that the weights are normally distributed find how many students weigh I) Between 130 and 148kg ii) more than 152kg

Solution:

Let μ be the mean and σ be the standard deviation. Then μ =140kg and σ =10pounds

(i) When x= 138, $z = \frac{x-\mu}{\sigma} = \frac{138-140}{10} = -0.2 = z_1$

When x= 138, $z = \frac{x-\mu}{\sigma} = \frac{148-140}{10} = 0.8 = z_2$

ii) When x=152, $\frac{x-\mu}{\sigma}$=$\frac{152-140}{10} = 1.2$=z_1

Therefore P(x>152)=P(z>z_1)=0.5-A(z_1)

=0.5-0.3849=0.1151

Therefore number of students whose weights are more than 152kg

=800x0.1151=92.

Exercise Problems:

1. Two coins are tossed simultaneously. Let X denotes the number of heads then find i) E(X) ii) E(X^2) iii)E(X^3) iv) V(X)
2. If f(x)=k$e^{-|x|}$ is probability density function in the interval, $-\infty < x < \infty$, then find i) k ii) Mean iii) Variance iv) P(0<x<4)

3. If X is a normal variate with mean 30 and standard deviation 5. Find the probabilities that i) $P(26 \leq X \leq 40)$ ii) $P(X \geq 45)$

4. The marks obtained in Statistics in a certain examination found to be normally distributed. If 15% of the students greater than or equal to 60 marks, 40% less than 30 marks. Find the mean and standard deviation.

MODULE– III
CORRELATION AND REGRESSION

MODULE III

CLOs	Course Learning Outcome
CLO 11	Explain multiple random variables and the covariance of two random variables.
CLO 12	Understand the concept of multiple random variables in real-world problems aspects of wireless communication system.
CLO 13	Calculate the correlation coefficient to the given data.
CLO 14	Contrast the correlation and regression to the real-world such as stock price and interest rates.
CLO 15	Calculate the regression to the given data.

CORRELATION

CORRELATION:

In a bivariate distribution, if the change in one variable effects the change in other variable, then the variables are called correlated.

Covariance: Covariance between two random variables X and Y is denoted by Cov(X,Y) is defined as E(XY)-E(X)E(Y)

If X and Y are independent then Cov(X,Y) = 0

Karl Pearson Correlation Coefficient

- Karl Pearson Correlation Coefficient between two r.v. X and Y usually denoted by r(X,Y) or simply r_{XY} is a numerical measure of a linear relationship between them and is defined as r = r(X,Y) = $\text{cov}(X,Y)/\sigma_x\sigma_y$

  It is also called product moment correlation coefficient.

- If (x_i,y_i); I = 1,2…n is bivariate distribution then, then

  Cov (X,Y) = E[{X-E(X)}{Y-E(Y)}]

  $$= (1/n)\sum(x_i-\bar{x})(y_i-\bar{y}) = (1/n)\sum x_iy_i - \bar{x}\,\bar{y}$$

$$\sigma_X^2 = E[X-E(X)]^2 = (1/n)\sum(x_i-\bar{x})^2 = (1/n)\sum x_i^2-(\bar{x})^2$$

$$\sigma_Y^2 = E[Y-E(Y)]^2 = (1/n)\sum(y_i-\bar{y})^2 = (1/n)\sum y_i^2-(\bar{y})^2$$

Computational formula for r(X,Y) = $\dfrac{\frac{1}{n}\sum xy - \overline{x}\overline{y}}{\sqrt{\left[\left(\frac{1}{n}\sum x^2\right) - \overline{x}^2\right]}\sqrt{\left[\left(\frac{1}{n}\sum y^2\right) - \overline{y}^2\right]}}$

-1 $\leq$ r $\leq$ 1

If r = 0 then X,Y are uncorrelated.

If r = -1 then correlation is perfect and negative.

If r = 1then the correlation is perfect and positive.

r is independent of change of origin and scale

Two independent variables are uncorrelated. Converse need not be true.

CORRELATION

The correlation coefficient for Bivariate frequency distribution:

The bivariate data on X on Y are presented in a two-way correlation table with n classes of Y placed along the horizontal lines and m classes of X along vertical lines and f_{ij} is the frequency of the individuals lying in i, j th cell.

$\sum_x f(x,y)$ =g(y),is the sum of the frequencies along any row and

$\sum_y f(x,y)$ =f(x),is the sum of the frequencies along any column.

$$\sum_x \sum_y f(x,y) = \sum_y \sum_x f(x,y) = \sum_x f(x) = \sum_y g(y) = \text{N}$$

$$\overline{\boldsymbol{x}} = \frac{1}{N}\sum_x xf(x), \qquad \bar{y} = \frac{1}{N}\sum_y yg(y)$$

$$\sigma_X^2 = \frac{1}{N}\sum_x x^2 f(x) - \bar{x}^2 \text{ and } \sigma_Y^2 = \frac{1}{N}\sum_y y^2 g(y) - \bar{y}^2$$

Cov(X,Y) = $\frac{1}{N}\sum_{x}\sum_{y} xyf(x,y)$ - $\bar{x}\,\bar{y}$

r= $\frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

CORRELATION PROBLEMS

1.Calculate the coefficient of correlation from the following data

x	12	9	8	10	11	13	7
y	14	8	6	9	11	12	13

Solution: Here $X = x - \bar{x}$ and $Y = y - \bar{y}$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{12+9+8+10+11+13+7}{7} = 10$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{14+8+6+9+11+12+13}{7} = 10.4$$

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	XY	X^2	Y^2
12	14	2	3.6	7.2	4	12.9
9	8	-1	-2.4	2.4	1	5.7
8	6	-2	-4.4	8.8	4	19.3
10	9	0	-1.4	0	0	1.9
11	11	1	0.6	0.6	1	0.3
13	12	3	1.6	4.8	9	2.5
7	13	-3	2.6	-7.8	9	6.7
				$\sum XY = 16$	$\sum X^2 = 28$	$\sum Y^2 = 49.3$

$\therefore$ Correlation Coefficient $$r = \frac{\sum XY}{\sqrt{\sum X^2 . \sum Y^2}}$$

$$= \frac{16}{\sqrt{28 \times 49.3}}$$

$$r = 0.43$$

r is positive.

problems

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

x	15	18	20	24	30	35	40	50
y	85	93	95	105	120	130	150	160

CORRELATION

- **Rank Correlation**: Let (x_i,y_i) for I = 1,2,…n be the ranks of the i^{th} individuals in the characteristics A and B respectively, Pearsonian coefficient of correlation between x_i and y_i are called rank correlation coefficient between A and B for that group of individual.

Spearman’s rank correlation

- **The Spearman’s rank correlation** between the two variables X and Y takes the values 1,2…n denoted by

  ρ and is defined as $\rho = 1 - \dfrac{\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$

  where $d_i = x_i - y_i$ (In general $x_i \neq y_i$)

Spearman’s rank correlation

In case, common ranks are given to repeated items, the common rank is the average of the ranks which these items would have assumed if they were slightly different from each other and the next item will get the rank next to the rank already assumed.

The adjustment or correction is made in the rank correlation formula. In the formula we add factor $\frac{m(m^2-1)}{12}$ to $\sum$ d^2, where m is the number of times an item is repeated. This correction factor is to be added to each repeated value in both X-series and Y- series.

-1 $\leq \rho \leq$ 1

Problems:

The ranks of 16 students in Mathematics and Statistics are as follows (1,1),(2,10),(3,3),(4,4),(5,5),(6,7),(7,2),(8,6),(9,8),(10,11),(11,15),(12,9),(13,14),(14,12),(15,16),(16,13). Calculate the rank correlation coefficient for proficiencies of this group in mathematics and statistics.

.

Rank Correlation Coefficient

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

$$= 1 - \frac{6 \times 136}{16 \times 225}$$

$$\therefore \rho = 0.8$$

REGRESSION

- Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.

The variable whose value is influenced or is to be predicted is called dependent variable and the variable, which influences the values or is used for the prediction is called independent variable.

REGRESSION

- Independent variable is also known as regressor or predictor or explanatory variable while the dependent variable is also known as regressed or explained variable.

If the variables in bivariate distributions are related we will find that the points in the scatter diagram will cluster round some curve called the “ curve of regression”. If the curve is a straight line, it is called line of regression

REGRESSION

- In the bivariate distribution (x_i,y_i) ; i = 1,2,….n Y is dependent variable and X is independent variable. The line of regression Y on X is Y = a + b X.

  i.e. Y-$\bar{y}$ = r$\dfrac{\sigma_Y}{\sigma_X}(X - \bar{x})$

  Similarly the line of regression X on Y is X = a + b Y

  i.e.,X- $\bar{x}$ = r$\dfrac{\sigma_X}{\sigma_Y}(Y - \bar{y})$

REGRESSION

If X and Y are any random variables the two regression lines are

$$Y - E(Y) = \frac{Cov(X,Y)}{\sigma_X^2}[X - E(X)]$$

$$X - E(X) = \frac{Cov(X,Y)}{\sigma_Y^2}[Y - E(Y)]$$

REGRESSION

- Both lines of regression passes through the point $\left(\bar{x},\bar{y}\right)$ i.e., the mean values $\left(\bar{x},\bar{y}\right)$ can be obtained at the point of intersection of regression lines.
- The slope of regression line Y on X is also called the regression coefficient Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X. We write, b_{YX} = Regression coefficient Y on X $= r\frac{\sigma_Y}{\sigma_X}$

REGRESSION

- The modulus value of the arithmetic mean of regression coefficient is not less than modulus value of correlation coefficient r.
- Regression coefficients are independent of the change of origin but not scale.
- If θ is the acute angle between two lines of regression then

$$\theta = \mathrm{Tan}^{-1}\left\{\frac{1-r^2}{|r|}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}\right)\right\}$$

REGRESSION

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable*and is referred to as Y. The variable we are basing our predictions on is called the *predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

REGRESSION

The example data in Table 1 are plotted in Figure 1. You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

REGRESSION

Lines of Regression:

- Regression coefficients are independent of the change of origin but not scale.
- If θ is the acute angle between two lines of regression then

$$\theta = \mathrm{Tan}^{-1}\left\{\frac{1-r^2}{|r|}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}\right)\right\}$$

REGRESSION

If   r = 0 then variables X and Y are uncorrelated. The lines of regressions are Y = $\bar{y}$ and X = $\bar{x}$ which are perpendicular to each other and are parallel to x- axis and y-axis respectively.

If r = ± 1 , the two lines of regression coincide.

REGRESSION

Problems:

1. Determine the regression equation which best fit to the following data:

x	10	12	13	16	17	20	25
y	10	22	24	27	29	33	37

Solution: The regression equation of y on x is $y = a + bx$

The normal equations are

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

REGRESSION

x	y	x^2	xy
10	10	100	100
12	22	144	264
13	24	169	312
16	27	256	432
17	29	289	493
20	33	400	660
25	37	625	925
$\sum x = 113$	$\sum y = 182$	$\sum x^2 = 1983$	$\sum xy = 3186$

REGRESSION

Substitute the above values in normal equations

$$\sum y = na + b\sum x \Rightarrow 182 = 7a + 113b ------(1)$$

$$\sum xy = a\sum x + b\sum x^2 \Rightarrow 3186 = 113a + 1983b ------(2)$$

REGRESSION

Solve equations (1) and (2) we get

$a = 0.7985$ and $b = 1.5611$

Now substitute a, b values in regression equation

$\therefore$ The regression equation of y on x is $y = 0.7985 + 1.5611x$

REGRESSION

Exercise Problems:

1. If θ is the angle between two regression lines and S.D. of Y is twice the S.D. of X and r=0.25,
2. Determine the regression equation which best fit to the following data:

x	10	12	13	16	17	20	25
y	10	22	24	27	29	33	37

Multiple correlation and Regression

- The coefficient of **multiple correlation** is a measure of how well a given variable can be predicted using a linear function of a set of other variables. It is the **correlation** between the variable's values and the best predictions that can be computed linearly from the predictive variables.

- **Multiple Regression** analysis is an extension of (simple) regression analysis in which two or more independent variables are used to estimate the value of dependent variable.

Least square regression planes fitting of N data points (X_1,X_2,X_3) in a three dimensional scatter diagram. The least square regression plane of X_1 on X_2 and X_3 is $X_1 = a + b X_2+cX_3$

simultaneously the normal equations:

$$\sum X_1 = an + b\sum X_2 + c\sum X_3$$

$$\sum X_1 X_2 = a\sum X_2 + b\sum X_2^2 + c\sum X_2 X_3$$

$$\sum X_1 X_3 = a\sum X_3 + b\sum X_2 X_3 + c\sum X_3^2$$

Similarly for the regression plane of X_2 on X_1 and X_3 and the regression plane of X_3 on X_1 and X_2

The linear regression equation of X_1 on X_2 ,X_3 and

X_4 can be written as

$X_1 = a + b\, X_2 + c\, X_3 + d\, X_4$

Problems:

1. Give the following data compute multiple coefficient of correlation of X_3 on X_1 and X_2.

X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2
X_3	90	72	54	42	30	12

Solution: here $n = 6, \quad \overline{X}_1 = \frac{48}{6} = 8, \quad \overline{X}_2 = \frac{42}{6} = 7, \quad \overline{X}_3 = \frac{300}{6} = 50$

Now we calculate values of r_{12}, r_{13} and r_{23}

Problems:

1. Give the following data compute multiple coefficient of correlation of X_3 on X_1 and X_2.

X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2
X_3	90	72	54	42	30	12

Solution: here $n = 6, \quad \overline{X}_1 = \frac{48}{6} = 8, \quad \overline{X}_2 = \frac{42}{6} = 7, \quad \overline{X}_3 = \frac{300}{6} = 50$

Now we calculate values of r_{12}, r_{13} and r_{23}

	$x_1 = X_1 - \bar{X}_1$			$x_2 = X_2 - \bar{X}_2$			$x_3 = X_3 - \bar{X}_3$					
S.NO	X_1	x_1	$x_1^{\ 2}$	X_2	x_2	x_2^2	X_3	x_3	$x_3^{\ 2}$	x_1x_2	x_2x_3	x_3x_1
1	3	-5	25	16	9	81	90	40	1600	-45	360	-200
2	5	-3	9	10	3	9	72	22	484	-9	66	-66
3	6	-2	4	7	0	0	54	4	16	0	0	-8
4	8	0	0	4	-3	9	42	-8	64	0	24	0
5	12	4	16	3	-4	16	30	-20	400	-16	80	-80
6	14	6	36	2	-5	25	12	-38	1444	-30	190	-228

$$r_{12} = \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2 \sum x_2^2}} = \frac{-100}{\sqrt{90 \times 140}} = -0.89$$

$$r_{12} = \frac{\sum x_1 x_3}{\sqrt{\sum x_1^2 \sum x_3^2}} = \frac{-582}{\sqrt{90 \times 4008}} = -0.97$$

$$r_{12} = \frac{\sum x_2 x_3}{\sqrt{\sum x_2^2 \sum x_3^2}} = \frac{720}{\sqrt{140 \times 4008}} = 0.96$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{13}r_{23}r_{12}}{1 - r_{12}^2}} = 0.987$$

MODULE– IV
TEST OF HYPOTHESIS-I

CLOs	Course Learning Outcome
CLO 16	Discuss the concept of sampling distribution of statistics and in particular describe the behavior of the sample mean.
CLO 17	Understand the foundation for hypothesis testing
CLO 18	Summarize the concept of hypothesis testing in real-world problem to selecting the best means to stop smoking.
CLO 19	Apply testing of hypothesis to predict the significance difference in the sample means
CLO 20	Apply testing of hypothesis to predict the significance difference in the sample proportions

SAMPLING

Sampling Distribution

- **Population** is the set or collection or totality of the objects, animate or inanimate, actual or hypothetical under study. Thus, mainly population consists of set of numbers measurements or observations, which are of interest.
- Size of the population N is the number of objects or observations in the population.
- Population may be finite or infinite.
- A finite sub-set of the population is known as **Sample.** Size of the sample is denoted by n.
- **Sampling** is the process of drawing the samples from a given population.
- If $n \geq 30$ the sampling is said to be **large sampling.**
- If $n < 30$ then the sampling is said to be **Small sampling.**

.

SAMPLING

- Population f(x) is a population whose probability distribution is f(x).If f(x) is binomial, Poisson or normal then the corresponding population is known as Binomial Population, Poisson population or normal Population.
- Samples must be representative of the population, sampling should be random.
- Random Sampling is one in which each member of the population has equal chances or probability of being included in the sample.
- Sampling where each member of a population may be chosen, more than once is called **Sampling with replacement**. A finite population, which is sampled with replacement, can theoretically be considered infinite since samples of any size can be drawn with out exhausting the population. For most practical purpose sampling from a finite population, which is very large, can be considered as sampling from an infinite population.
- If each member cannot be chosen more than once it is called sampling with out replacement.

SAMPLING

- Random samples (Finite population): A set of observations X_1, X_2......X_n, constitute a random sample of size n from a finite population of size N, if its values are chosen so that each subset of n of the N elements of the population has same probability if being selected.
- Random sample (Infinite Population): A set of observations X_1, X_2......X_n constitute a random sample of size n from infinite population f(x) if:

  (i) Each X_i is a r.v. whose distribution is given by f(x)

  (ii) These n r.v.'s are independent
- **Sample Mean** X_1, X_2......X_n is a random sample of size n the sample mean is a r.v. defined by $\overline{X} = \frac{X_1 + X_2 + ...X_n}{n}$

$$\sigma_Y^2 = E[\ Y\text{-}E(Y)]^2 = (1/n)\sum(y_i - \bar{y})^2 = (1/n)\sum y_i^2 - (\bar{y})^2$$

SAMPLING

- **Sample Variance** $X_1, X_2......X_n$ is a random sample of size n the sample variance is a r.v. defined by $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$ and is a measure of variability of data about the mean.
- **Sample Standard deviation** is the positive square root of the sample variance.
- **Degrees of freedom (d o f)** of a statistic is the positive integer denoted by υ, equals to n-k where n is the number of independent observations of the random sample and k is the number of population parameters which are calculated using sample data. Thus **d o f** $\upsilon = n – k$ is the difference between n, the sample size and k, the number of independent constraints imposed on the observations in the sample.

- **Sampling Distributions:** The probability distribution of a sample statistic is often called as sampling distribution of the statistic.
- The standard deviation of the sampling distribution of a statistic is called **Standard Error(S.E)**
- The mean of the sampling distribution of means, denoted by $\mu_{\bar{x}}$, is given by $E(\bar{X}) = \mu_{\bar{x}} = \mu$ where μ is the mean of the population.
- If a population is infinite or if sampling is with replacement, then the variance of the sampling distribution of means, denoted by $\sigma_{\bar{x}}^2$ is given by $E[(\bar{X} - \mu)^2] = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ where σ^2 is the variance of the population.
- If the population is of siqe N, if sampling is without replacement, and if the sample size is $n \leq N$ then $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

CENTRAL LIMIT THEOREM

- Central limit theorem: If $\overline{x}$ is the mean of a sample of size n taken from a population having the mean μ and the finite variance σ^2, then Z=$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is a r.v. whose distribution function approaches that of the standard normal distribution as n→∞
- If $\overline{x}$ is the mean of a sample of size n taken from a finite population of size N with mean μ and variance σ^2 then Z=$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$ is a r.v whose distribution function approaches that of the standard normal distribution as n→∞
- The normal distribution provides an excellent approximation to the sampling distribution of the mean $\overline{x}$ for n as small as 25 or 30
- If the random samples come from a normal population, the sampling distribution of the mean is normal regardless of the size of the sample

TESTING OF HYPOTHESIS

- Statistical decisions are decisions or conclusions about the population parameters on the basis of a random sample from the population.
- Statistical hypothesis is an assumption or conjecture or guess about the parameters of the population distribution

- **Null Hypothesis (N.H)** denoted by H_0 is statistical hypothesis, which is to be actually tested for acceptance or rejection. NH is the hypothesis, which is tested for possible rejection under the assumption that it is true.
- Any Hypothesis which is complimentary to the N.H is called an **Alternative Hypothesis** denoted by H_1
- Simple Hypothesis is a statistical Hypothesis which completely specifies an exact parameter. N.H is always simple hypothesis stated as a equality specifying an exact value of the parameter. E.g. N.H = $H_0 : \mu = \mu_0$ N.H. = $H_0 : \mu_1 - \mu_2 = \delta$
- Composite Hypothesis is stated in terms of several possible values.
- Alternative Hypothesis(A.H) is a composite hypothesis involving statements expressed as inequalities such as < , > or $\neq$

  i) A.H : $H_1 : \mu > \mu_0$ (Right tailed) ii) A.H : $H_1 : \mu < \mu_0$ (Left tailed)

  iii) A.H : $H_1 : \mu \neq \mu_0$ (Two tailed alternative)

ERRORS IN SAMPLING

- **Errors in sampling:**

  Type I error: Reject H_0 when it is true

  Type II error: Accept $\mathbf{H_0}$ when it is wrong (i.e) accept if when H_1 is true.

	Accept H_0	Reject H_0
H_0 is True	Correct Decision	Type 1 error
H_0 is False	Type 2 error	Correct Decision

- If P{ Reject H_0 when it is true}= P{ Reject H_0 | H_0}=α and P{ Accept H_0 when it is false}= P{ Accept H_0 | H_1} = β then α,β are called the sizes of Type I error and Type II error respectively. In practice, type I error amounts to rejecting a lot when it is good and type II error may be regarded as accepting the lot when it is bad.

- α and β are referred to as producers risk and consumers risk respectively.
- A region (corresponding to a statistic t) in the sample space S that amounts to rejection of H_0 is called critical region of rejection.
- Level of significance is the size of the type I error (or maximum producer's risk)
- The levels of significance usually employed in testing of hypothesis are 5% and 1% and is always fixed in advance before collecting the test information.
- A test of any statistical hypothesis where AH is one tailed(right tailed or left tailed) is called a **one–tailed test.** If AH is two-tailed such as: H_0**:** $\mu = \mu_0$, against the AH. H_1 **:** $\mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$) is called **Two-Tailed Test.**

- The value of test statistics which separates the critical (or rejection) region and the acceptance region is called **Critical value or Significant value**. It depends upon (i) The level of significance used and (ii) The Alternative Hypothesis, whether it is two-tailed or single tailed

Critical Value (Z_α)	Level of significance (α)		
	1%	5%	10%
Two-Tailed test	$-Z_{\alpha/2} = -2.58$ $Z_{\alpha/2} = 2.58$	$-Z_{\alpha/2} = -1.96$ $Z_{\alpha/2} = 1.96$	$-Z_{\alpha/2} = -1.645$ $Z_{\alpha/2} = 1.645$
Right-Tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left-Tailed Test	$-Z_\alpha = -2.33$	$- Z_\alpha = -1.645$	$-Z_\alpha = -1.28$

- When the size of the sample is increased, the probability of committing both types of error I and II (i.e) α and β are small, the test procedure is good one giving good chance of making the correct decision.
- P-value is the lowest level (of significance) at which observed value of the test statistic is significant.
- A test of Hypothesis (T. O.H) consists of
  1. Null Hypothesis (NH) : H_0
  2. Alternative Hypothesis (AH) : H_1
  3. Level of significance: α
  4. Critical Region pre determined by α
  5. Calculation of test statistic based on the sample data.
  6. Decision to reject NH or to accept it.

CONFIDENCE INTERVALS

Maximum error E of a population mean μ by using large sample mean is

E = $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

The most widely used values for 1-α are 0.95 and 0.99 and the corresponding values of $Z_{\alpha/2}$ are $Z_{0.025}$ = 1.96 and $Z_{0.005}$ = 2.575

Sample size n = $\left[Z_{\alpha/2}\frac{\sigma}{E}\right]^2$

CONFIDENCE INTERVAL

- Confidence interval for μ (for large samples n ≥ 30) σ known

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- If the sampling is without replacement from a population of finite size N then the confidence interval for μ with known is

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} < \mu < \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

1. Large sample confidence interval for μ - σ unknown is

   $$\bar{x} - Z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{S}{\sqrt{n}}$$

   Large sample confidence interval for μ_1 - μ_2 (where σ_1 and σ_2 are unknowns)

   $$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}$$

   The end points of the confidence interval are called **Confidence Limits.**

PROBLEMS

PROBLEMS: **1.** 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

i) The mean of the population

ii) The standard deviation of the population

iii) The mean of the sampling distribution of means

iv) The standard deviation of the sampling distribution of means

Solution: Given that N=5, n=2 and

i. Mean of the population

$$\mu = \sum \frac{x_i}{N} = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6$$

ii. Variance of the population

$$\sigma^2 = \sum \frac{(x_i - \bar{x})^2}{N} = \frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5}$$

$$= \frac{16+9+0+4+25}{5}$$

$$= 10.8$$

$$\sigma = 3.29$$

Sampling with replacement(infinite population):
The total number of samples with replacement is

$$N^n = 5^2 = 25$$

There 25 samples can be drawn

$$\left\{\begin{matrix} (2,2) & (2,3) & (2,6) & (2,8) & (2.11) \\ (3,2) & (3,3) & (3,6) & (3,8) & (3,11) \\ (6,2) & (6,3) & (6.6) & (6,8) & (6,11) \\ (8,2) & (8,3) & (8,6) & (8,8) & (8,11) \\ (11,2) & \grave{}(11,3) & (11,6) & (11,8) & (11,11) \end{matrix}\right\}$$

The sample means are

$$\left\{\begin{matrix} 2 & 2.5 & 4 & 5 & 6.5 \\ 2.5 & 3 & 4.5 & 5.5 & 7 \\ 4 & 4.5 & 6 & 7.0 & 8.5 \\ 5 & 5.5 & 7 & 8 & 9.5 \\ 6.5 & 7 & 8.5 & 9.5 & 11 \end{matrix}\right\}$$

iii)The mean of the sampling distribution of means is

$$\mu_{\bar{x}} = \frac{2+2.5+4+5+6.5+----+11}{25}$$

=6

iv)The standard deviation of the sampling distribution of means

$$\sigma_{\bar{x}}^2 = \frac{(2-6)^2+(2.5-6)^2+----+(11-6)^2}{25}$$

$= 5.40$

$$\sigma_{\bar{x}} = 2.32$$

2. A population consists of five numbers4, 8, 12, 16, 20, 24. Consider all possible samples of size two which can be drawn without replacement from this population. Find

i) The mean of the population
ii) The standard deviation of the population
iii) The mean of the sampling distribution of means
iv)the standard deviation of the sampling distribution of means

Solution: Given that N=6, n=2 and

i. Mean of the population

$$\mu = \sum \frac{x_i}{N} = \frac{4+8+12+16+20+24}{6} = \frac{84}{6} = 14$$

ii. Variance of the population

$$\sigma^2 = \sum \frac{(x_i - \bar{x})^2}{N} = \frac{(4-14)^2 + (8-14)^2 + (12-14)^2 + (16-14)^2 + (20-14)^2 + (24-14)^2}{6}$$

$$= \frac{100+36+4+4+36+100}{6}$$

$$= 46.67$$

$$\sigma = 3.29$$

Sampling without replacement (finite population):

The total number of samples without replacement is $N_{c_n} = 6_{c_2} = 15$

There 15 samples can be drawn

$$\left\{\begin{matrix} (4,8) & (4,12) & (4,16) & (4,20) & (4,24) \\ (8,12) & (8,16) & (8,20) & (8,24) & \\ (12,16) & (12,20) & (12,24) & & \\ (16,20) & (16,24) & & & \\ (20,24) & \text{`} & & & \end{matrix}\right\}$$

The sample means are

$$\left\{\begin{matrix} 6 & 8 & 10 & 12 & 14 \\ 10 & 12 & 14 & 16 & \\ 14 & 16 & 18 & & \\ 18 & 20 & & & \\ 22 & & & & \end{matrix}\right\}$$

iii)The mean of the sampling distribution of means is

$$\mu_{\bar{x}} = \frac{6+8+10+12-----20+22}{15}$$

=14

iv)The standard deviation of the sampling distribution of means

$$\sigma_{\bar{x}}^{2} = \frac{(6-14)^{2}+(8-14)^{2}+----+(22-14)^{2}}{15}$$

=18.67

$$\sigma_{\bar{x}} = 4.32$$

PROBLEMS

3. 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 mean of the population (μ) = 155 cm

Standard deviation of the population (σ) = 15 cm

Sample size (n) = 36

Mean of sample ($\bar{x}$) = 157 cm

Now $Z = \dfrac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

$$= \frac{157 - 155}{\frac{15}{\sqrt{36}}}$$

=0.8

$\therefore P(\bar{x} \leq 157) = P(Z < 0.8)$

= $0.5 + P(0 \leq Z \leq 0.8)$

=0.5+0.2881

$\therefore P(\bar{x} \leq 157)$ = 0.7881

The mean of the sampling distribution of means is

$$\mu_{\bar{x}} = \frac{2+2.5+4+5+6.5+----+11}{25}$$

=6

The standard deviation of the sampling distribution of means

$$\sigma_{\bar{x}}^2 = \frac{(2-6)^2+(2.5-6)^2+----+(11-6)^2}{25}$$

$$= 5.40$$

$$\sigma_{\bar{x}} = 2.32$$

4)Determine a 95% confidence interval for the mean of normal distribution with variance 0.25, using a sample of size 100 values with mean 212.3.

Solution: Given Sample size (n) = 100

Standard deviation of sample (σ) = $\sqrt{0.25} = 0.5$

Mean of sample ($\bar{x}$) = 212.3 and $Z_{\alpha/2}$ = 1.96(for 95%)

$\therefore$ Confidence interval = $\left(\bar{x} - Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}\right)$

$= \left(212.3 - 1.96.\frac{0.5}{\sqrt{100}}, 212.3 + 1.96.\frac{0.5}{\sqrt{100}}\right)$

= (212.202, 212.398)

Exercise Problems:

1. Samples of size 2 are taken from the population 1, 2, 3, 4, 5, 6. Which can be drawn without replacement? Find

i) The mean of the population

ii) The standard deviation of the population

iii) The mean of the sampling distribution of means

iv) The standard deviation of the sampling distribution of means

2. If a 1-gallon can of paint covers on an average 513 square feet with a standard deviation of 3   1.5 square feet, what is the probability that the mean area covered by a sample of 40 of

these 1-gallon cans will be anywhere from 510to 520 square feet?

3. What is the size of the smallest sample required to estimate an unknown proportion to within a maximum error of 0.06 with at least 95% confidence.

4. A random sample of 400 items is found to have mean 82 and standard deviation of 18. Find the maximum error of estimation at 95% confidence interval. Find the confidence limits for the mean if $\bar{x}$=82.

5. A sample of size 300 was taken whose variance is 225 and mean is 54. Construct 95% confidence interval for the mean.

LARGE SAMPLE TESTS

Test statistic for T.O.H. in several cases are

Statistic for test concerning mean σ known

Z = $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$

Statistic for large sample test concerning mean with σ unknown

Z = $\frac{\overline{X}-\mu_0}{S/\sqrt{n}}$

Statistic for test concerning difference between the means

Z = $\frac{(\overline{X}_1 - \overline{X}_2) - \delta}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$ under NH $H_0 : \mu_1 - \mu_2 = \delta$ against the AH, $H_1 : \mu_1 - \mu_2 > \delta$

or $H_1 : \mu_1 - \mu_2 < \delta$ or $H_1 : \mu_1 - \mu_2 \neq \delta$

Statistic for large samples concerning the difference between two means (σ_1 and σ_2 are unknown)

Z = $\frac{(\overline{X}_1 - \overline{X}_2) - \delta}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$

Statistics for large sample test concerning one proportion

Z = $\frac{X - np_o}{\sqrt{np_0(1-p_0)}}$ under the N.H: H_0**:** p = p_o against H_1: p $\neq$ p_0 or p > p_0 or

p <P_0

Statistic for test concerning the difference between two proportions

Z= $\frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$ with $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$ under the NH : H_0: p_1=p_2 against the AH

H_1:$p_1 < p_2$ or $p_1 > p_2$ or $p_1 \neq p_2$

- Large sample confidence interval for difference of two proportions (p_1- p_2) is

$$\left(\frac{x_1}{n_1}-\frac{x_2}{n_2}\right)\pm Z_{\alpha/2}\sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1}+\frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}}$$

- Maximum error of estimate E = $Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ with observed value x/n substituted for p we obtain an estimate of E
- Sample size n = $p(1-p)\left(\frac{Z_{\alpha/2}}{E}\right)^2$ when p is known

  n= $\frac{1}{4}\left(\frac{Z_{\alpha/2}}{E}\right)^2$ when p is unknown

- One sided confidence interval is of the form p < $(1/2n)\chi_\alpha^2$ with (2n+1) degrees of freedom.

LARGE SAMPLE TESTS PROBLEMS

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

Solution: Given n=400, $\bar{x}=40$ and μ=38 and σ=10

1. **Null hypothesis(H_0):** μ=38
2. **Alternative hypothesis(H_1):** $\mu \neq 38$
3. **Level of significance:** α=0.05 and z_α=1.96
4. **Test statistic:** $z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

$$z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}=\frac{40-38}{\frac{10}{\sqrt{400}}}=4$$

$$|z|=4$$

Confidence interval = $\left(\bar{x} - Z_{\alpha}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$

$$=\left(40 - 1.96\frac{10}{\sqrt{400}}, 40 + 1.96\frac{10}{\sqrt{400}}\right)$$

Samples of students were drawn from two universities and from their weights in kilograms mean and S.D are calculated and shown below make a large sample test to the significance of difference between means.

	MEAN	S.D	SAMPLE SIZE
University-A	55	10	400
University-B	57	15	100

Solution: *Given n_1=400, n_2=100, $\bar{x}_1$=55, $\bar{x}_2$=57*
S_1=10 and S_2=15

1. ***Null hypothesis(H_0):*** $\bar{x}_1 = \bar{x}_2$
2. ***Alternative hypothesis(H_1):*** $\bar{x}_1 \neq \bar{x}_2$
3. ***Level of significance:*** α*=0.05 and* Z_α*=1.96*
4. ***Test statistic:*** $Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ = $\frac{55-57}{\sqrt{\frac{100}{400} + \frac{225}{100}}}$*=-1.26*

   $|Z| =$ *1.26*

5. ***Conclusion:***

$\therefore$ $|Z| < Z_\alpha$

$\therefore$We accept the Null hypothesis.

LARGE SAMPLE TESTS PROBLEMS

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

Solution: Given n=400, $\bar{x} = 40$ and μ=38 and σ=10

1. **Null hypothesis(H_0):** μ=38
2. **Alternative hypothesis(H_1):** $\mu \neq$38
3. **Level of significance:** α=0.05 and z_α=1.96
4. **Test statistic:** $Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{40-38}{\frac{10}{\sqrt{400}}}$=4 $|Z| =$4

5. **Conclusion:**

$\therefore$ $|Z| > z_\alpha$

$\therefore$We reject the Null hypothesis.

Confidence interval = $\left(\bar{x} - Z_\alpha \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right)$

= $\left(40 - 1.96 \frac{10}{\sqrt{400}}, 40 + 1.96 \frac{10}{\sqrt{400}}\right)$

= $(39.02, 40.98)$

1. Samples of students were drawn from two universities and from their weights in kilograms mean and S.D are calculated and shown below make a large sample test to the significance of difference between means.

	MEAN	S.D	SAMPLE SIZE
University-A	55	10	400
University-B	57	15	100

Solution: Given n_1=400, n_2=100, $\bar{x}_1$=55, $\bar{x}_2$=57

S_1=10 and S_2=15

1. **Null hypothesis(H_0):** $\bar{x}_1$**=**$\bar{x}_2$
2. **Alternative hypothesis(H_1):** $\bar{x}_1 \neq \bar{x}_2$
3. **Level of significance:** α**=**0.05 and Z_α=1.96
4. **Test statistic:** $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ = $\dfrac{55-57}{\sqrt{\dfrac{100}{400} + \dfrac{225}{100}}}$ **=-**1.26

$|Z| =$ 1.26

5. **Conclusion:**

$\therefore$ $|Z| < Z_\alpha$

$\therefore$ We accept the Null hypothesis.

1. **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% level of significance?**

Solution: Given n=400, x=540

$p = \frac{x}{n} = \frac{540}{1000} = 0.54$

$P = \frac{1}{2} = 0.5$, $Q = 0.5$

1. **Null hypothesis(H_0):** P=0.5
2. **Alternative hypothesis(H_1):** $P \neq 0.5$
3. **Level of significance:** α=1% and z_α=2.58
4. **Test statistic:** $Z = \frac{P-p}{\sqrt{\frac{PQ}{n}}}$

$Z = \frac{P-p}{\sqrt{\frac{PQ}{n}}} = \frac{0.54-0.5}{\sqrt{\frac{0.5\times0.5}{1000}}} = 2.532$ $|Z| = 2.532$ $\therefore$ $|Z| < z_\alpha$ $\therefore$ We accept the Null hypothesis.

4.Random sample of 400 men and 600 women were asked whether they would like to have flyover near their residence .200 men and 325 women were in favour of proposal. Test the hypothesis that the proportion of men and women in favour of proposal are same at 5% level.

Solution: Given n_1=400, n_2=600 , $x_1 = 200$ and $x_2 = 325$

$$p_1 = \frac{200}{400} = 0.5$$

$$p_2 = \frac{325}{600} = 0.541$$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{400 \times \frac{200}{400} + 600 \times \frac{325}{600}}{400 + 600} = 0.525$$

1. **Null hypothesis(H_0):** $p_1 = p_2$
2. **Alternative hypothesis(H_1):** $p_1 \neq p_2$
3. **Level of significance:** α=0.05 and z_α=1.96
4. **Test statistic:** $Z = \dfrac{p_1 - p_2}{\sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ $=$ $\dfrac{0.5 - 0.541}{\sqrt{0.525 \times 0.425\left(\dfrac{1}{400} + \dfrac{1}{600}\right)}} = -1.28$

$|Z| =$1.28

5. **Conclusion:**

$\therefore$ $|Z| < z_\alpha$

$\therefore$ We accept the Null hypothesis.

MODULE– V

TEST OF HYPOTHESIS-II

CLOs	Course Learning Outcome
CLO 21	Use Student t-test to predict the difference in sample means.
CLO 22	Apply F-test to predict the difference in sample variances.
CLO 23	Understand the characteristics between the samples using Chi-square test.

t-distribution

- If $\overline{X}$ is the mean of a random sample of size n taken from a normal population having the mean μ and the variance σ^2 , and $s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ then t=$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$ is a r.v. having the t- distribution with the parameter υ = (n-1)dof

- The overall shape of a t-distribution is similar to that of a normal distribution both are bell shaped and symmetrical about the mean. Like the standard normal distribution t-distribution has the mean 0, but its variance depends on the parameter υ (nu), called the number of degrees of freedom. The variance of t- distribution exceeds1, but it approaches 1 as n→∞. The t-distribution with υ-degree of freedom approaches the standard normal distribution as υ→∞.

  The standard normal distribution provides a good approximation to the t-

  distribution for samples of size 30 or more

Producer of ‘gutkha’ claims that the nicotine content in his ‘gutkha’ on the average is 83 mg. can this claim be accepted if a random sample of 8 ‘gutkhas’ of this type have the nicotine contents of 2.0,1.7,2.1,1.9,2.2,2.1,2.0,1.6 mg.

Solution: Given n=8 and μ=1.83 mg

1. Null hypothesis(H_0): μ=1.83
2. Alternative hypothesis(H_1): $\mu \neq$ 1.83
3. Level of significance: α=0.05

t_α for n-1 degrees of freedom $t_{0.05}$ for 8-1 degrees of freedom is 1.895

Test statistic:

$$t = \frac{\overline{x} - \mu}{\frac{S}{\sqrt{n}}}$$

x	$(x-\bar{x})$	$(x-\bar{x})^2$
2.0	0.05	0.0025
1.7	-0.25	0.0625
2.1	0.15	0.0225
1.9	-0.05	0.0025
2.2	0.25	0.0625
2.1	0.15	0.0225
2.0	0.05	0.0025
1.6	-0.35	0.1225
Total=15.6		

$\overline{x} = \frac{15.6}{8} = 1.95$ and $S^2 = \sum \frac{(x-\overline{x})^2}{n-1} = \frac{0.3}{7}$ S=0.21

$t = \frac{\overline{x}-\mu}{\frac{S}{\sqrt{n}}} = \frac{1.95-1.83}{\frac{0.21}{\sqrt{8}}}$ =1.62 $|t| = 1.62$

Conclusion:

$\therefore |t| < t_{\alpha}$ $\therefore$ We accept the Null hypothesis.

The means of two random samples of sizes 9,7 are 196.42 and 198.82.the sum of squares of deviations from their respective means are 26.94,18.73.can the samples be considered to have been the same population?

Solution: Given n_1=9, n_2=7, $\bar{x}_1$=196.42, $\bar{x}_2$=198.82 and

$\sum(x_i-\bar{x}_1)^2$=26.94,

$\sum(x_i-\bar{x}_2)^2$=18.73

$\therefore S^2=\dfrac{\sum(x_i-x_1)^2+\sum(x_i-x_2)^2}{n_1+n_2-2}$=3.26

$\Rightarrow$S=1.81

Null hypothesis(H_0): $\bar{x}_1 = \bar{x}_2$

Alternative hypothesis(H_1): $\bar{x}_1 \neq \bar{x}_2$

Level of significance: α=0.05

t_α for n_1+n_2-2 degrees of freedom

$t_{0.05}$ for 9+7-2=14 degrees of freedom is 2.15

Test statistic: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ = $\dfrac{196.42-198.82}{(1.81)\sqrt{\dfrac{1}{9}+\dfrac{1}{7}}}$ **=-2.63** $|t| =$ 2.63

Conclusion:

$\therefore$ $|t| > t_\alpha$ $\therefore$ We reject the Null hypothesis.

F-distribution

- If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 respectively, taken from two normal populations having the same variance, then $F=\frac{S_1^2}{S_2^2}$ is a r.v.

  having the F- distribution with the parameter's $\upsilon_1=n_1-1$ and $\upsilon_2=n_2-1$ are called the numerator and denominator degrees of freedom respectively.
- $F_{1-\alpha}(\upsilon_1,\upsilon_2)=\frac{1}{F_\alpha(\upsilon_2,\upsilon_1)}$

In one sample of 8 observations the sum of squares of deviations of the sample values from the sample mean was 84.4 and another sample of 10 observations it was 102.6 .test whether there is any significant difference between two sample variances at at 5% level of significance.

Solution: Given n_1=8, n_2=10, $\sum(x_i - \bar{x}_1)^2$=84.4 and $\sum(x_i - \bar{x}_2)^2$=102.6

$$S_1{}^2 = \frac{\sum(x_i - x_1)^2}{n_1 - 1} = \frac{84.4}{7} = 12.057$$

$$S_2{}^2 = \frac{\sum(x_i - x_1)^2}{n_2 - 1} = \frac{102.6}{9} = 11.4$$

Null hypothesis(H_0): ${S_1}^2 = {S_2}^2$

Alternative hypothesis(H_1): ${S_1}^2 \neq {S_2}^2$

Level of significance: α=0.05

F_α For $(n_1 - 1, n_2 - 1)$ degrees of freedom

$F_{0.05}$ For (7,9) degrees of freedom is 3.29

Test statistic: $F = \frac{S_1^2}{S_2^2}$ $= \frac{12.057}{11.4}$ =1.057

$|F| = 1.057$

Conclusion: $\therefore$ $|F| < F_\alpha$ $\therefore$ We accept the Null hypothesis.

Chi-square test

- If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2 , then $\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$ is a r.v. having the chi-square distribution with the parameter υ = n-1
- The chi-square distribution is not symmetrical

The following table gives the classification of 100 workers according to gender and nature of work. Test whether the nature of work is independent of the gender of the worker.

	Stable	Unstable	Total
Male	40	20	60
Female	10	30	40
Total	50	50	100

Solution: Given that

$$\text{Expected frequencies} = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

$\frac{90\times100}{200}$=45	$\frac{90\times100}{200}$=45	90
$\frac{90\times100}{200}$=55	$\frac{90\times100}{200}$=55	110
100	100	200

Calculation of χ^2 :

Observed Frequency(O_i)	Expected Frequency(E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
60	45	225	5
30	45	225	5
40	55	225	4.09
70	55	225	4.09
			18.18

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 18.18$$

1. Null hypothesis(H_0): $O_i = E_i$

2. Alternative hypothesis(H_1): $O_i \neq E_i$

3. Level of significance: $\alpha = 0.05$

$\chi_\alpha{}^2$ For (r-1)(c-1) degrees of freedom

$\chi_{0.05}{}^2$ For (2-1)(2-1)=1 degrees of freedom is 3.84

$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ =18.18

1. Null hypothesis(H_0): $\mathrm{O_i = E_i}$
2. Alternative hypothesis(H_1): $\mathrm{O_i \neq E_i}$
3. Level of significance: α=0.05

   $\chi_\alpha{}^2$ For (r-1)(c-1) degrees of freedom

   $\chi_{0.05}{}^2$ For (2-1)(2-1)=1 degrees of freedom is 3.84
4. Test statistic**:** $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ =18.18,

$|\chi^2| = 1.057$

Conclusion: $\therefore$ $|\chi^2|$ > $\chi_\alpha{}^2$

$\therefore$We reject the Null hypothesis.

Thank you