SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – I – SMTA1308 –MATHEMATICAL STATISTICS

UNIT I CONCEPT OF SAMPLE SPACE

Events, Definition of Probability (Classical, Statistical & Axiomatic) - Addition and Multiplication laws of probability -Independence - Conditional Probability - Bayes' theorem - Simple problems.

Random Experiment

An experiment whose outcome or result can be predicted with certainty is called a Deterministic experiment.

Although all possible outcomes of an experiment may be known in advance the outcome of a particular performance of the experiment cannot be predicted owing to several unknown causes. Such an experiment is called a random experiment.

Example:

Whenever a fair dice is thrown, it is known that any of the 6 possible outcomes will occur, but it cannot be predicted what exactly the outcome will be.

Sample Space

The set of all possible outcomes of an experiment. They are assumed equally likely.

Event

A sub-set of S consisting of possible outcomes.

Probability Definitions

Mathematical Definition

If A is an event of a random experiment and S be the sample space, then

P(A) = (No. of ways A can occur)/(Total no. of possible outcomes)

i.e $P(A) = \frac{n(A)}{n(S)}$

Axiomatic Definition

Given a random experiment, Let A be a mutually, exclusive and exhaustive likely event and S be the sample space of all possible events. The number P(A) is called the probability measure of A or simply the probability of A, If the following properties (axioms of probability) are satisfied

Axiom 1: $p(A) \geq 0$

Axiom 2: $p(S) = 1$

Axiom 3: for a countable collection of mutually exclusive events $A_1, A_2, \ldots A_n$ in S,

$$p(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n) = p(A_1) + p(A_2) + p(A_3) + \cdots p(A_n)$$

Complement of an Event

If A is an event then its complement is A^c or A' or $\bar{A}$

A complement of an event A can be stated as that which does NOT contain the occurrence of A.

The probability of the complement of an event is denoted as $P(A^c)$ or P(A').

$P(A^c) = 1 - P(A)$

or it can be stated, $P(A)+P(A^c) = 1$

Theorems

Theorem 1: The probability of an impossible event is zero.

Proof: Let S be sample space (certain events) and ϕ be the impossible event.

Certain events and impossible events are mutually exclusive.
$P(S \cup \phi) = P(S) + P(\phi)$ (Axiom 3)
$S \cup \phi = S$
$P(S) = P(S) + P(\phi)$
$P(\phi) = 0$, hence the result.

Theorem 2: If $\overline{A}$ is the complementary event of A, $P(\overline{A}) = 1 - P(A) \leq 1$.

Proof: Let A be the occurrence of the event
$\overline{A}$ be the non-occurrence of the event .
Occurrence and non-occurrence of the event are mutually exclusive.

$$P(A \cup \overline{A}) = P(A) + P(\overline{A})$$
$$A \cup \overline{A} = S \quad \Rightarrow \quad P(A \cup \overline{A}) = P(S) = 1$$
$$\therefore \quad 1 = P(A) + P(\overline{A})$$
$$P(\overline{A}) = 1 - P(A) \leq 1.$$

Theorem 3: Addition Theorem On Probability

If A and B be any two events then probability of A or B is
$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$

Proof: We know, $A = A\overline{B} \cup AB$ and $B = \overline{A}B \cup AB$

$$\therefore \quad P(A) = P(A\overline{B}) + P(AB) \text{ and } P(B) = P(\overline{A}B) + P(AB) \qquad \text{(Axiom 3)}$$
$$P(A) + P(B) = P(A\overline{B}) + P(AB) + P(\overline{A}B) + P(AB)$$
$$= P(A \cup B) + P(A \cap B)$$

Note: 1 P(A or B) is the probability of the occurrence of at least one of the events.
Note 2: P(A and B) is the probability of the occurrence of both A and B at the same time.

Mutually Exclusive Events:

Mutually exclusive events are those where the occurrence of one indicates the non-occurrence of the other
or When two events cannot occur at the same time, they are considered mutually exclusive.

Note: For a mutually exclusive event, P(A and B) = 0.

Problems

1. A coin is thrown 3 times .what is the probability that at least one head is obtained?
   Sol: Sample space = [HHH, HHT, HTH, THH, TTH, THT, HTT, TTT]
   Total number of ways = 2 × 2 × 2 = 8. Fav. Cases = 7
   P (A) = 7/8
   or
   P (of getting at least one head) = 1 – P (no head)⇒ 1 – (1/8) = 7/8

2. Find the probability of getting a numbered card when a card is drawn from the pack of 52 cards.
   Sol: Total Cards = 52. Numbered Cards = (2, 3, 4, 5, 6, 7, 8, 9, 10) 9 from each suit 4 × 9 = 36
   P (E) = 36/52 = 9/13

3. There are 5 green 7 red balls. Two balls are selected one by one without replacement. Find the probability that first is green and second is red.
   Sol: P (G) × P (R) = (5/12) x (7/11) = 35/132

4. Find the probability that a leap year has 52 Sundays.
   Sol: A leap year can have 52 Sundays or 53 Sundays. In a leap year, there are 366 days out of which there are 52 complete weeks & remaining 2 days. Now, these two days can be (Sat, Sun) (Sun, Mon) (Mon, Tue) (Tue, Wed) (Wed, Thur) (Thur, Friday) (Friday, Sat).
   So there are total 7 cases out of which (Sat, Sun) (Sun, Mon) are two favorable cases.
   So, P (53 Sundays) = 2 / 7
   Now, P(52 Sundays) + P(53 Sundays) = 1
   So, P (52 Sundays) = 1 - P(53 Sundays) = 1 – (2/7) = (5/7)

5. From a pack of cards, three cards are drawn at random. Find the probability that each card is from different suit.
   Sol: Total number of cases = $^{52}C_3$
   One card each should be selected from a different suit. The three suits can be chosen in $^{4}C_3$ ways
   The cards can be selected in a total of ($^{4}C_3$) x ($^{13}C_1$) x ($^{13}C_1$) x ($^{13}C_1$)
   Probability = $^{4}C_3$ x ($^{13}C_1$)^3 / $^{52}C_3$
   = 4 x (13)^3 / $^{52}C_3$
6. Two dice are thrown together. What is the probability that the number obtained on one of the dice is multiple of number obtained on the other dice?
   Sol: Total number of cases = 6^2 = 36
   Since the number on a die should be multiple of the other, the possibilities are
   (1, 1) (2, 2) (3, 3) ------ (6, 6) --- 6 ways
   (2, 1) (1, 2) (1, 4) (4, 1) (1, 3) (3, 1) (1, 5) (5, 1) (6, 1) (1, 6) --- 10 ways
   (2, 4) (4, 2) (2, 6) (6, 2) (3, 6) (6, 3) -- 6 ways
   Favorable cases are = 6 + 10 + 6 = 22. So, P (A) = 22/36 = 11/18
7. A lot consists of 10 good articles, 4 with minor defects and 2 with major defects. Two articles are chosen from the lot at random(with out replacement). Find the probability that (i) both are good, (ii) both have major defects, (iii) at least 1 is good, (iv) at most 1 is good, (v) exactly 1 is good, (vi) neither has major defects and (vii) neither is good.

   Solution :

   (i) P(both are good) = $\frac{10C_2}{16C_2} = \frac{3}{8}$

   (ii) P(both have major defects) = $\frac{2C_2}{16C_2} = \frac{1}{120}$

   (iii) P(at least 1 is good) = $\frac{10C_1 6C_1 + 10C_2}{16C_2} = \frac{7}{8}$

   (iv) P(at most 1 is good) = $\frac{10C_0 6C_2 + 10C_1 6C_1}{16C_2} = \frac{5}{8}$

(v) P(exactly 1 is good) = $\frac{10C_1 6C_1}{16C_2} = \frac{1}{2}$

(vi) P(neither has major defects) = $\frac{14C_2}{16C_2} = \frac{91}{120}$

(vii) P(neither is good) = $\frac{6C_2}{16C_2} = \frac{1}{8}$.

Conditional probability

Conditional probability is calculating the probability of an event given that another event has already occurred

Definition of Conditional Probability:

The probability of an event A is given then another event B occurred is called conditional probability of A given B. It is denoted by P(A/B).

P(A/B) = P(A ∩ B)/P(B) **or P(A|B) = P (A and B) / P(B)**

Similarly, when the probability of Y given X is

P(B/A) = P(A ∩ B)/P(A)

Example: In a class, 40% of the students study math and science. 60% of the students study math. What is the probability of a student studying science given he/she is already studying math?

Solution

P(M and S) = 0.40

P(M) = 0.60

P(S|M) = P(M and S)/P(S) = 0.40/0.60 = 2/3 = 0.67

Multiplication Theorem of Probability:

In an experiment suppose, A and B are any two events then probabilities of both A and B is given by

P(A ∩ B) = P(A) · P(B/A) ---------- (i)

or

P(A ∩ B) = P(B) · P(A/B) ---------- (ii)

A theorem of Total Probability

Statement: Let $A_1, A_2, \ldots, A_n$ be a set of events associated with a sample space S, where all the events $A_1, A_2, \ldots, A_n$ have non-zero probability of occurrence and they form a partition of S. Let B be any event associated with S, then

$$P(B) = \sum_{k=1}^{n} P(A_k)P(B|A_k)$$

Baye's Theorem

Statement: Let $A_1, A_2, \ldots, A_n$ be a set of events associated with a sample space S, where all the events $A_1, A_2, \ldots, A_n$ have non-zero probability of occurrence and they form a partition of S. Let B be any event associated with S, then

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{\sum_{k=1}^{n} P(A_k)P(B|A_k)}$$

for any k = 1, 2, 3, …., n

Problems

1. Three bags contain 3 red, 7 black; 8 red, 2 black, and 4 red & 6 black balls respectively. 1 of the bags is selected at random and a ball is drawn from it. If the ball drawn is red, find the probability that it is drawn from the third bag.

   Sol: Let E_1, E_2, E_3 and A are the events defined as follows.
   E_1 = First bag is chosen
   E_2 = Second bag is chosen
   E_3 = Third bag is chosen
   A = Ball drawn is red
   Since there are three bags and one of the bags is chosen at random,

so $P(E_1) = P(E_2) = P(E_3) = 1 / 3$
If E_1 has already occurred, then first bag has been chosen which contains 3 red and 7 black balls. The probability of drawing 1 red ball from it is 3/10.
So, $P(A/E_1) = 3/10$,
similarly $P(A/E_2) = 8/10$, and $P(A/E_3) = 4/10$.
We are required to find $P(E_3/A)$ i.e. given that the ball drawn is red,
The probability that the ball is drawn from the third bag

$$P(E_3|A) = \frac{P(E_3)P(A|E_3)}{P(A|E_1) + P(E_2)P(A|E_2) + P(E_{23})P(A|E_3)}$$

(by Baye's rule)
$$= \frac{\frac{1}{3}\times\frac{4}{10}}{\frac{1}{3}\times\frac{3}{10}+\frac{1}{3}\times\frac{8}{10}+\frac{1}{3}\times\frac{4}{10}} = \frac{4}{15}.$$

2. **The bag I contains 4 white and 6 black balls while another Bag II contains 4 white and 3 black balls. One ball is drawn at random from one of the bags and it is found to be black. Find the probability that it was drawn from Bag I.**

Solution:
Let A_1 be the event of choosing the bag I, A_2 the event of choosing the bag II and A be the event of drawing a black ball.
Then, $P(A_1) = P(A_2) = 1/2$
Also, $P(A|E_1)$ = P(drawing a black ball from Bag I) = 6/10 = 3/5
$P(A|E_2)$ = P(drawing a black ball from Bag II) = 3/7
By using Bayes' theorem, the probability of drawing a black ball from bag I out of two bags,

$$P(E_1|A) = \frac{P(E_1)P(A|E_1)}{P(A \mid E_1) + P(E_2)P(A|E_2)}$$

$$= \frac{1/2 \times 3/5}{1/2 \times 3/7 + 1/2 \times 3/5} = 7/12$$

3. A bolt is manufactured by 3 machines A, B, C . A turn out twice as many items as B and machines B and C produce equal number of item. 2%of bolts produced by A and B are defective
   and4% of bolts produced by C are defective. All bolts are put into 1 stock pile and 1 is chosen from this pile. What is the probability that it is defective?
   Solution
   Let A= The event in which the item has been produced by machine A
   B= The event in which the item has been produced by machine B
   C= The event in which the item has been produced by machine C
   D= The event of the item being defective
   $P(A) = \frac{1}{2}, P(B) = \frac{1}{4} = P(C)$
   $P(D/A) = \frac{2}{100} = P\left(\frac{D}{B}\right), P(D/C) = \frac{4}{100}$
   By theorem on total probability
   $$P(D) = P(A)P(D/A) + P(B)P(D/B) + P(C)P(D/C)$$
   $$= \frac{1}{2} \times \frac{2}{100} + \frac{1}{4} \times \frac{2}{100} + \frac{1}{4} \times \frac{4}{100} = \frac{1}{40}$$

4. For a certain binary communication channel, the probability that a transmitted 0 is received as a 0 is 0.95and the probability that a transmitted 1is received as1 is 0.9. If the probability that a 0 is transmitted is 0.4. Find the probability that (i) a 1 is received and (ii) a 1 was transmitted given that a 1 was received.

Solution

Let A= The event of transmitting 1
$\bar{A}$= The event of transmitting 0
B= The event of receiving 1
$\bar{B}$= The event of receiving 0

$P(A) = 0.6, P(\bar{A}) = 0.4$
$P(B/A) = 0.9, (\bar{B}/\bar{A}) = 0.9, (B/\bar{A}) = 0.05$
By the theorem of total probability

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

By Baye's theorem,

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)) + P(\bar{A})P\left(\frac{B}{\bar{A}}\right)} = \frac{0.6 \times .9}{0.56} = \frac{27}{28}$$

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – II – SMAT1308 – MATHEMATICAL STATISTICS

UNIT II RANDOM VARIABLES

(Discrete and Continuous) Distribution function - Expected values & moments - Moment generation function -probability generation functions – Examples.

Definition: Random Variable
A random variable is a function that assigns a real number X(s) to every element s∈S, Where S is the sample space corresponding to a random experiment.

Discrete Random Variable
If X is a random variable that can take a finite number or countably infinite number of values, X is called a discrete RV, when the RV is discrete, the possible values of X may be assumed as $x_1, x_2, \ldots x_n, \ldots$.

Probability Function

The ***probability distribution*** of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function.

Suppose a random variable *X* may take *k* different values, with the probability that $X = x_i$ defined to be $P(X = x_i) = p_i$. Then p_i is said to be probability function or probability mass function or point probability function provided p_i must satisfy the following conditions:

(i) $0 \leq p_i \leq 1$ *for all i*

(ii) $p_1 + p_2 + \ldots + p_k = 1$*(i.e)* $\cdot \sum_i p_i = 1$

Example:
Suppose a variable X can take the values 1, 2, 3, or 4.
The probabilities associated with each outcome are described by the following table:

Outcome	1	2	3	4
Probability	0.1	0.3	0.4	0.2

The probability that *X* is equal to 2 or 3 is the sum of the two probabilities
$P(X = 2 \text{ or } X = 3) = P(X = 2) + P(X = 3) = 0.3 + 0.4 = 0.7$.
Similarly, the probability that *X* is greater than 1 is equal to
$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 1) = 1 - 0.1 = 0.9$.

Continuous Random Variable
If X is an RV that can take all values in an interval, then X is called a continuous random variable.

Probability Density Function
If X is a continuous random variable such that f(x) is called the probability density function of X provided f(x) satisfies the following conditions:
(i) $f(x) \geq o$, for all $x \in R_X$
(ii) $\int_{R_X} f(x)dx = 1$
Moreover, $P(a \leq X \leq b) or\ P(a < X < b) = \int_a^b f(x)dx$
Cumulative Distribution Function (CDF),
If X is an RV, discrete or continuous, then $P(X \leq x)$ is called the cumulative distribution function of X or distribution function of X and denoted as F(x).
If X is discrete,

$$F(x) = \sum_j x_j$$

If X is continuous,

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

Properties of the cdf F(x)

1. F(x) is a non-decreasing function of x_i
   (i.e) If $x_1<x_2$ then $F(x_1)\leq F(x_2)$.
2. $F(-\infty)=0$ and $F(\infty)=1$.
3. If X is a discrete R V taking values $x_1, x_2, ...,$
   where $x_1< x_2< x_3<....<x_{i-1}<x_i<...$, then $P(X=x_i)=F(x_i)-F(x_{i-1})$.
4. If X is a continuous RV, then $\frac{d}{dx}F(x) = f(x), at\ all\ points\ where\ F(x)\ is\ differentiable.$

Problems:

1. A random variable X has the following probability distribution

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

Find (i) the value of K, (ii) P(1.5<X<4.5/X>2)
(iii) The smallest vale of λ for which P(X≤λ)>1/2.

Solution
(i) We know that for a discrete RV X,

$$\sum_i p_i = 1$$

Therefore, $10K^2+9K=1$
$10K^2+9K-1=0$
(10K-1)(K+1)=0Æ
K=1/10 or K=-1.
But If K=-1, then the value of P(x) is negative.

Therefore K=1/10
(ii) P(1.5<X<4.5/X>2)=P(A/B) $= \frac{P(A\cap B)}{P(B)}$

$$\frac{P(1.5<X<4.5)\cap P(X>2)}{P(X>2)} = \frac{P(X=3)+P(X=4)}{P(X=3)+P(X=4)+P(X=5)+P(X=6)+P(X=7)}$$

$$= \frac{\frac{5}{10}}{\frac{7}{10}} = \frac{5}{7}$$

(iii) By trails,

$$P(X \leq 0) = 0; P(X \leq 2) = \frac{3}{10}$$

$$P(X \leq 3) = \frac{5}{10}; P(X \leq 4) = \frac{8}{10}$$

Therefore, the smallest value of λ satisfying the condition P (X ≤ λ)>1/2 is 4.

2. If the random variable X takes the values 1,2, 3 and 4 such that 2P(X=1)=3P(X=2)=P(X=3)=5P(X=4). Find the probability distribution and cumulative distribution function of X.

solution

Let P(X=3)=30K.

Therefore 2P(x=1)=30K

$$\Rightarrow P(x=1)=\frac{30K}{2}=15K$$

$$P(X=2)=10$$

$$P(X=4)=6K$$

$$\because \sum P_i=1$$

$$\therefore 15K+10K+30K+6K=1$$

$$\Rightarrow K=\frac{1}{61}$$

The proobability distribution of X is given in the following table

X=i	1	2	3	4
P_i	15/61	10/61	30/61	6/61

The cdf F(x) is defined a F(x)=P(X≤x)

When x<1, F(x)=o

When 1≤x<2, F(x)=P(X=1)=15/61

When 2≤x<3, F(x)=P(X=1)+P(X=2)=25/61

When 2≤x<4, F(x)= P(X=1)+P(X=2)+P(X=3)=55/61

When X≥4, F(x)= P(X=1)+P(X=2)+P(X=3)+P(X=4)=1

3. A continuous RV has a pdf $f(x)=3x^2$, 0≤x≤1.Find a and b such that

(i) P(X≤a)=P(X>a) , and (ii) P(X>b)=0.05

Solution

(i) P(X≤a)=P(X>a

$$\int_0^a 3x^2dx=\int_a^1 3x^2dx$$

$$\Rightarrow a^3=1-a^3$$

$$\Rightarrow 2a^3=1\Rightarrow a^3=0.5$$

$$\therefore a=0.7937$$

$$P(X>b)=0.05$$

$$\int_b^1 3x^2dx=0.05$$

$$b^3=95$$

$$b=0.9830$$

4. If the density function of a continuous RVX is given by

f(x)=ax 0≤x≤1

=a , 1≤x≤2

=3a-ax, 2≤x≤3

=0, elsewhere

(i)Find the value of a

(ii)Find the cdf of X

Solution

(i) Since f(x) is a pdf,

$$\int_{R_X} f(x)dx = 1$$

$$\Rightarrow \int_0^3 f(x)d = 1$$

$$\int_0^1 f(x)dx + \int_1^2 f(x)dx + \int_2^3 f(x)dx = 1$$

$$\int_0^1 axdx + \int_1^2 adx + \int_2^3 (3a - ax)dx = 1$$

$$2a = 1$$

$$a = 0.5$$

$$F(x) = P(X \le x) = 0, when x < 0$$

$$F(x) = \int_0^x \frac{x}{2}dx = x^4, when 0 \le x \le 1$$

$$= \int_0^1 \frac{x}{2}dx + \int_1^x \frac{1}{2}dx = \frac{x}{2} - \frac{1}{2} when 1 \le x \le 2$$

$$= \int_0^1 \frac{x}{2}dx + \int_1^2 \frac{1}{2}dx + \int_2^x \left(\frac{3}{2} - \frac{x}{2}\right)dx = \frac{3}{2}x - \frac{x^2}{4} - \frac{5}{4}, when 2 \le x \le 3$$

$$F(x) = P(X \le x) = 0, when x < 0$$

$$F(x) = \int_0^x \frac{x}{2}dx = x^4, when 0 \le x \le 1$$

$$= \int_0^1 \frac{x}{2}dx + \int_1^x \frac{1}{2}dx = \frac{x}{2} - \frac{1}{2} when 1 \le x \le 2$$

$$= \int_0^1 \frac{x}{2}dx + \int_1^2 \frac{1}{2}dx + \int_2^x \left(\frac{3}{2} - \frac{x}{2}\right)dx = \frac{3}{2}x - \frac{x^2}{4} - \frac{5}{4}, when 2 \le x \le 3$$

5.

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

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

Solution:

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

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

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

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

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

$$K\pi = 1$$

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

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

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

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

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

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

Expectation

For a discrete Rv

The expected value is the sum of: [(each of the possible outcomes) × (the probability of the outcome occurring)].

$$E(X) = \sum_i x_i P_i$$

$$E(X^2) = \sum_i x_i^{\ 2} P_i$$

Properties of Expectations

1. E(ax+bY)=aE(X)+bE(Y)
2. E(aX)=aE(X)
3. E(aX+b)=aE(X)+ b
4. EXY)= E(X)E(Y), If X and Y are independent RV

Example

What is the expected value when we roll a fair die?

There are six possible outcomes: 1, 2, 3, 4, 5, 6. Each of these has a probability of 1/6 of occurring. Let X represent the outcome of the experiment.

Therefore P(X = 1) = 1/6 (this means that the probability that the outcome of the experiment is 1 is 1/6)

P(X = 2) = 1/6 (the probability that you throw a 2 is 1/6)
P(X = 3) = 1/6 (the probability that you throw a 3 is 1/6)
P(X = 4) = 1/6 (the probability that you throw a 4 is 1/6)
P(X = 5) = 1/6 (the probability that you throw a 5 is 1/6)
P(X = 6) = 1/6 (the probability that you throw a 6 is 1/6)

$E(X) = 1\times P(X = 1) + 2\times P(X = 2) + 3\times P(X = 3) + 4\times P(X=4) + 5\times P(X=5) + 6\times P(X=6)$

Therefore $E(X) = 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 = \underline{7/2}$

So the expectation is 3.5 . If you think about it, 3.5 is halfway between the possible values the die can take and so this is what you should have expected.

Variance

The variance of a random variable tells us something about the spread of the possible values of the variable. For a discrete random variable X, the variance of X is written as Var(X).

- $Var(X) = E[(X - m)^2]$ where m is the expected value E(X)

This can also be written as:

- $Var(X) = E(X^2) - m^2 = E(X^2)-(E(X))^2$

The *standard deviation* of X is the square root of Var(X).

Note that the variance does not behave in the same way as the expectation when we multiply and add constants to random variables. In fact:

- $Var[aX + b] = a^2Var(X)$

Because: $Var[aX + b] = E[(aX + b)^2] - (E [aX + b])^2$.

$= E[a^2X^2 + 2abX + b^2] - (aE(X) + b)^2$
$= a^2E(X^2) + 2abE(X) + b^2 - a^2E^2(X) - 2abE(X) - b^2$
$= a^2E(X^2) - a^2E^2(X) = a^2Var(X)$

Properties of Variance

1. $Var(X)\geq 0$
2. $Var(aX+b)=a^2Var(X)$
3. $Var(aX+bY)= a^2Var(X)+ b^2Var(Y)$, If X and Y are independent

Definition: Let X be a **continuous** random variable with p.d.f. $f_X(x)$. The expected value of X is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

For Var(X), we use

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 .$$

Now

$$\begin{aligned} E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_1^2 x^2 \times 2x^{-2}\,dx &= \int_1^2 2\,dx \\ &= \Big[2x\Big]_1^2 \\ &= 2 \times 2 - 2 \times 1 \\ &= 2. \end{aligned}$$

Thus

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 \\ &= 2 - \{2\log(2)\}^2 \\ &= 0.0782. \end{aligned}$$

Example: Let X be a continuous random variable with p.d.f.

$$f_X(x) = \begin{cases} 2x^{-2} & \text{for } 1 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{E}(X)$ and $\mathrm{Var}(X)$.

$$\begin{aligned} E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\,dx = \int_1^2 x \times 2x^{-2}\,dx &= \int_1^2 2x^{-1}\,dx \\ &= \Big[2\log(x)\Big]_1^2 \\ &= 2\log(2) - 2\log(1) \\ &= 2\log(2). \end{aligned}$$

Moments

The ***n*th moment** of a distribution (or set of data) about a number is the expected value of the *n*th power of the deviations about that number. Moments are about the mean,
and about the origin.

- The *n*th moment of a distribution about the origin is given by $E(X^n)$
- The *n*th moment of a distribution about the mean is given by $E((X-\mu))^n$

Then each type of measure includes a moment definition.

- The expected value, E(X), is the first moment about the origin.
- The variance, Var(X), is the second moment about the mean, $E((X-\mu)^2$
- A common definition of skewness is the third moment about the mean, $E((X-\mu)^3$
- A common definition of kurtosis is the fourth moment about the mean, $E((X-\mu)^4$

Since moments about the origin are typically much easier to compute than moments about the mean, alternative formulas are often provided.

- $Var(X)=E((X-\mu)^2=E(X^2)-E(X)^2$
- $Var(X)=E((X-\mu)^2=E(X^2)-E(X)^2$. This formula was derived earlier.
- $Skew(X)=E((X-\mu)^3=E(X^3)-3E(X)E(X^2)+2E(X)^3$
- $Kurt(X)=E((X-\mu)^4)=E(X^4)-4E(X)E(X^3)+6E(X)2E(X^2)-3E(X)^4$

Moment Generating Functions

Since each moment is an expected value, and the definition of expected value involves either a sum (in the discrete case) or an integral (in the continuous case), it would seem that the computation of moments could be tedious. However, there is a single expected value function whose derivatives can produce each of the required moments. This function is called a **moment generating function**.
In particular, if X is a random variable, and either P(x) , f(x) are the PMF and PDF of the distribution (the first is discrete, the second continuous), then the moment generating function is defined by the following formulas.

Problem
Find moment generating function to the following probability distribution

$X = x$	$P(X = x)$
$X = 0$	0.4
$X = 1$	0.35
$X = 2$	0.25

Solution

Moment generating function

$$\begin{aligned} M_X(t) &= E(e^{tx}) \\ &= \sum_{\text{all } x} e^{tx} P(x) \\ &= 0.4e^{0t} + 0.35e^{1t} + 0.25e^{2t} \\ &= 0.4 + 0.35e^{t} + 0.25e^{2t} \end{aligned}$$

Now the moments can be obtained by differentiating $M_X(t)$

$$M_X'(t) = 0.35e^t + 0.5e^{2t} \qquad M_X'(0) = 0.35 + 0.5 = 0.85$$
$$M_X''(t) = 0.35e^t + e^{2t} \qquad M_X''(t) = 0.35 + 1 = 1.35$$
$$M_X^{(3)}(t) = 0.35e^t + 2e^{2t} \qquad M_X^{(3)}(t) = 0.35 + 2 = 2.35$$
$$M_X^{(4)}(t) = 0.35e^t + 4e^{2t} \qquad M_X^{(4)}(t) = 0.35 + 4 = 4.35$$

Find the moment generating function for $f(x) = 4e^{-4x}$, $0 \le x < \infty$

$$\begin{aligned} M_X(t) &= E(e^{tx}) \\ &= \int_0^\infty e^{tx} f(x)\,\mathrm{d}x \\ &= \int_0^\infty e^{tx} 4e^{-4x}\,\mathrm{d}x \\ &= \int_0^\infty 4e^{(t-4)x}\,\mathrm{d}x \\ &= \frac{4}{t-4} e^{(t-4)x}\Big|_0^\infty \qquad , \ t<4 \text{(assumed)} \\ &= \frac{-4}{t-4} = -4(t-4)^{-1} \end{aligned}$$

The moments are given below by differentiating the moment generating function $M_X(t)$

$$M_X'(t) = 4(t-4)^{-2} \qquad M_X'(0) = 4(-4)^{-2} = \frac{1}{4}$$
$$M_X''(t) = -8(t-4)^{-3} \qquad M_X''(t) = -8(-4)^{-3} = \frac{1}{8}$$
$$M_X^{(3)}(t) = 24(t-4)^{-4} \qquad M_X^{(3)}(t) = 24(-4)^{-4} = \frac{3}{32}$$
$$M_X^{(4)}(t) = -96(t-4)^{-5} \qquad M_X^{(4)}(t) = -96(-4)^{-5} = \frac{3}{32}$$

Other measures are

$$\begin{aligned} E(X) &= M'(0) = \frac{1}{4} \\ Var(X) &= E(X^2) - E(X)^2 \\ &= \frac{1}{8} - \left(\frac{1}{4}\right)^2 = \frac{1}{16} \\ Skew(X) &= E(X^3) - 3E(X)E(X^2) + 2E(X)^3 \\ &= \frac{3}{32} - 3\left(\frac{1}{4}\right)\left(\frac{1}{8}\right) + 2\left(\frac{1}{4}\right)^3 = \frac{1}{32} \\ Kurt(X) &= E(X^4) - 4E(X)E(X^3) + 6E(X)^2E(X^2) - 3E(X)^4 \\ &= \frac{3}{32} - 4\left(\frac{1}{4}\right)\left(\frac{3}{32}\right) + 6\left(\frac{1}{4}\right)^2\left(\frac{1}{8}\right) - 3\left(\frac{1}{4}\right)^4 = \frac{9}{256} \end{aligned}$$

Probability Generating Function (PGF)

The probability generating function (PGF) is a useful tool for dealing with discrete random variables taking values 0, 1, 2, ... Its particular strength is that it gives us an easy way of characterizing the distribution of X +Y when X and Y are independent. In general it is difficult to find the distribution

of a sum using the traditional probability function. The PGF transforms a sum into a product and enables it to be handled much more easily.

The name probability generating function also gives us another clue to the role of the PGF. The PGF can be used to generate all the probabilities of the distribution. This is generally tedious and is not often an efficient way of calculating probabilities. However, the fact that it can be done demonstrates that the PGF tells us everything there is to know about the distribution.

Definition: Let X be a discrete random variable taking values in the non-negative integers $\{0,1,2,\ldots\}$. The **probability generating function (PGF)** of X is $G_X(s) = \mathbb{E}(s^X)$, *for all* $s \in \mathbb{R}$ *for which the sum converges.*

Calculating the probability generating function

$$G_X(s) = \mathbb{E}\left(s^X\right) = \sum_{x=0}^{\infty} s^x \mathbb{P}(X = x).$$

Properties of PGF

1. $G_X(0) = \mathbb{P}(X = 0)$:

$$\begin{aligned} G_X(0) &= 0^0 \times \mathbb{P}(X=0) + 0^1 \times \mathbb{P}(X=1) + 0^2 \times \mathbb{P}(X=2) + \ldots \\ \therefore \quad G_X(0) &= \mathbb{P}(X=0). \end{aligned}$$

2. $G_X(1) = 1$: $\quad G_X(1) = \sum_{x=0}^{\infty} 1^x \mathbb{P}(X=x) = \sum_{x=0}^{\infty} \mathbb{P}(X=x) = 1.$

PGF of some Discrete distributions

Example 1 For Binomial Distribution

Let $X \sim \text{Binomial}(n,p)$, so $\mathbb{P}(X=x) = \binom{n}{x} p^x q^{n-x}$ for $x = 0, 1, \ldots$

$$\begin{aligned} G_X(s) &= \sum_{x=0}^{n} s^x \binom{n}{x} p^x q^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (ps)^x q^{n-x} \\ &= (ps+q)^n \quad \textit{by the Binomial Theorem: true for a} \end{aligned}$$

Thus $G_X(s) = (ps+q)^n$ *for all* $s \in \mathbb{R}$.

Example 2: **Poisson Distribution**

Let $X \sim \text{Poisson}(\lambda)$, so $\mathbb{P}(X = x) = \dfrac{\lambda^x}{x!}e^{-\lambda}$ for $x = 0, 1, 2, \ldots$.

$$G_X(s) = \sum_{x=0}^{\infty} s^x \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda s)^x}{x!}$$

$$= e^{-\lambda} e^{(\lambda s)} \quad \textit{for all } s \in \mathbb{R}.$$

Thus $G_X(s) = e^{\lambda(s-1)}$ *for all* $s \in \mathbb{R}$.

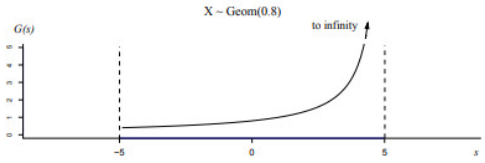
Example 3: **Geometric Distribution**

Let $X \sim \text{Geometric}(p)$, so $\mathbb{P}(X = x) = p(1-p)^x = pq^x$ for $x = 0, 1, 2, \ldots$, where $q = 1 - p$.

$$G_X(s) = \sum_{x=0}^{\infty} s^x p q^x$$

$$= p \sum_{x=0}^{\infty} (qs)^x$$

$$= \frac{p}{1 - qs} \quad \textit{for all } s \textit{ such that } |qs| < 1.$$



Thus $G_X(s) = \dfrac{p}{1 - qs}$ *for* $|s| < \dfrac{1}{q}$.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – III–SMTA1308 –Mathematical Statistsics

UNIT III CONCEPTS OF BIVARIATE DISTRIBUTION

Correlation

The word correlation is used in everyday life to denote some form of association. In statistical terms, we use correlation to denote association between two quantitative variables. Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

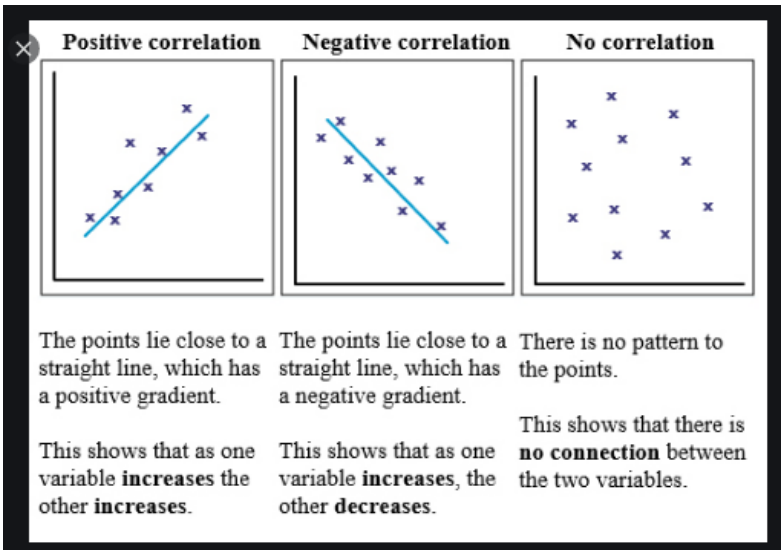
Correlation Coefficient

The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two variables. The correlation coefficient is scaled so that it is always between -1 and +1. When r is close to 0 this means that there is little relationship between the variables and the farther away from 0, r is, in either the positive or negative direction, the greater the relationship between the two variables.

Types of Correlation

- Positive Correlation – when the value of one variable increases with respect to another.
- Negative Correlation – when the value of one variable decreases with respect to another.
- No Correlation – when there is no linear dependence or no relation between the two variables.
- Correlation can have the following

  1 is a perfect positive correlation
  0 is no correlation (the values don't seem linked at all)
  -1 is a perfect negative correlation



Correlation

Correlation shows the relation between two variables. The correlation coefficient shows the measure of correlation. To compare two data sets we use a measure called correlation coefficient.

Karl-Pearson Correlation Coefficient Formula

The most common formula is the Pearson Correlation coefficient used for linear dependency between the data set. The value of the coefficient lies between -1 to +1. When the coefficient comes

down to zero, then the data is considered not related. While, if we get the value of +1, then the data are positively correlated and -1 has a negative correlation.

Karl-Pearson correlation coefficient is given by

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

$$r(X,Y) = \frac{\dfrac{\sum_{i=1}^{n} x_i y_i}{n} - (\bar{x})(\bar{y})}{\sqrt{\dfrac{\sum_{i=1}^{n} x_i^2}{n} - (\bar{x})^2}\ \sqrt{\dfrac{\sum_{i=1}^{n} y_i^2}{n} - (\bar{y})^2}}$$

$$r(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 . \sum_{i=1}^{n} (y_i - \bar{y})^2\right]^{\frac{1}{2}}}$$

The coefficient of multiple correlation lies between 0 and 1.

Problems

1.

Calculate the correlation coefficient between X and Y from the following data:

$$\sum_{i=1}^{15} (X_i - \bar{X})^2 = 136 \quad \sum_{i=1}^{15} (Y_i - \bar{Y})^2 = 138 \qquad \sum_{i=1}^{15} (X_i - \bar{X})(Y_i - \bar{Y}) = 122$$

Solution:

We have $r(X,Y) = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 . \sum_{i=1}^{n} (y_i - \bar{y})^2\right]^{\frac{1}{2}}} = \dfrac{122}{\sqrt{136}\sqrt{138}}$ r(X,Y) = 0.89

The correlation coefficient is a dimensionless number; it has no units of measurement. The maximum value r can achieve is 1, and its minimum value is –1. Therefore, for any given set of observations, $-1 \le r \le 1$.

2. A computer operator while calculating the coefficient of correlation between two variables X and Y for 25 pairs of observations obtained the following constants: $\sum X = 125$, $\sum Y = 100$, $\sum XY = 508$, $\sum X^2 = 650$, $\sum Y^2 = 460$. However it was later discovered at the time of checking that he had copied two pairs as (6,14) and (8,6) while the correct pairs were (8,12) and (6,8). Obtain the correct correlation coefficient.

Solution:

The formula involved with the given data is, $$r(X,Y) = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n} - (\bar{x})(\bar{y})}{\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - (\bar{x})^2}\ \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n} - (\bar{y})^2}}$$

The Corrected $\sum X$ = Incorrect $\sum X$ – (6+8) + (8+6) = 125

Corrected $\sum Y$ = Incorrect $\sum Y$ – (14+6) + (12+8) = 100

Corrected $\sum X^2$ = Incorrect $\sum X^2$ – $(6^2+8^2) + (8^2+6^2) = 650$

Corrected $\sum Y^2$ = Incorrect $\sum Y^2$ – $(14^2+6^2) + (12^2+8^2) = 436$

Corrected $\sum XY$ = Incorrect $\sum XY$ – (84+48) + (96+48) = 520

Now the correct value of correlation coefficient is, $$r(X,Y) = \frac{\frac{520}{25} - (5 \times 4)}{\sqrt{\frac{650}{25} - 5^2}\ \sqrt{\frac{436}{25} - 4^2}} = 0.67$$

3. Calculate the correlation coefficient for the following heights (in inches) of fathers x and their sons Y.

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

Solution

X	Y	XY	X^2	Y^2
65	67	4355	4225	4489
66	68	4488	4359	4624
67	65	4355	4489	4285
68	72	4896	4624	5184
69	72	4968	4761	5184
70	69	4830	4900	4761
72	71	5112	5184	5041
$\sum X = 544$	$\sum Y = 552$	$\sum XY = 37560$	$\sum X^2 = 37028$	$\sum Y^2 = 38132$

$$\overline{X} = \frac{\sum x}{n} = \frac{544}{8} = 68$$

$$\overline{Y} = \frac{\sum y}{n} = \frac{552}{8} = 69$$

$$\overline{XY} = 68 \times 69 = 4692$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum x^2 - \overline{X}^2} = \sqrt{\frac{1}{8}(37028) - 68^2} = \sqrt{4628.5 - 4624} = 2.121$$

$$Cov(X,Y) = \frac{1}{n}\sum XY - \overline{X}\,\overline{Y} = \frac{1}{8}(37650) - 68 \times 69$$

$$= 4695 - 4692 = 3$$

The correlation coefficient of X and Y is given by

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{3}{(2.121)(2.345)}$$

$$= \frac{3}{4.973} = 0.6032\,.$$

4. Compute the correlaion between X and Y , using the following data

X	1	3	5	7	8	10
Y	8	12	15	17	18	20

Solution

x_i	y_i	X_i^2	Y_i^2	$X_i Y_i$
1	8	1	64	8
3	12	9	144	36
5	15	25	225	75
7	17	49	289	119
8	18	64	324	144
10	20	100	400	200
7	17	49	289	119
34	90	248	1446	582

Thus $n_1 = 6$

$$\sum_i x_i = 34, \sum_i y_i = 90, \sum_i x_i^{\,2} = 248, \sum_i y_i^{\,2} = 1446,$$

$$\sum_i x_i y_i = 582$$

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}}$$

$$r = 0.9879$$

5. The data below summarized the relationship between the number of employees (x) and the number of openings (y) at 11 Boston area hospitals.

(b) The coefficient of determination, $r^2 = 0.8444 = 0.713$

This means that 71% of the variations in the number of openings can be explained by the linear relationship between it and the number of employees.

(b) Find the coefficient of determination and interpret its value.

Solution: n = 11

(a) The correlation coefficient is given by the formula: $r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2} \cdot \sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$

So from data: $r = \frac{11(18267023) - (56562)(2611)}{\sqrt{11(456525234) - (56562)^2} \cdot \sqrt{11(818149) - (2611)^2}}$

So $r = \frac{5325387.1}{(42690.9561) \cdot (1477.2671)} = 0.8444$

properties of correlation

- The coefficient of **Correlation** lies between -1 and +1.
- Coefficients of **Correlation** are independent of Change of Origin.
- Coefficients of **Correlation** possess the **property** of symmetry.
- Coefficient of **Correlation** is independent of Change of Scale.
- Co-efficient of **correlation** measures only linear **correlation** between X and Y

Rank correlation

If the ranks of two variables X and Y are given(ranks are not repeated)

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i = difference in paired ranks and n = number of cases.

For tied ranks

$CF = m(m^2-1)/12$

$$\rho = \frac{1 - 6\sum d^2}{n(n^2-1)} + CF$$

Problems

Example 1: Suppose we have ranks of 8 students of B.Sc. in Statistics and Mathematics. On the basis of rank we would like to know that to what extent the knowledge of the student in Statistics and Mathematics is related.

Rank in Statistics	1	2	3	4	5	6	7	8
Rank in Mathematics	2	4	1	5	3	8	7	6

Solution: Spearman's rank correlation coefficient formula is

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Let us denote the rank of students in Statistics by R_x and rank in Mathematics by R_y. For the calculation of rank correlation coefficient we have to find $\sum_{i=1}^{n} d_i^2$ which is obtained through the following table:

Rank in Statistics (R_x)	**Rank in Mathematics** (R_y)	**Difference of Ranks** $(d_i = R_x - R_y)$	d_i^2
1	2	−1	1
2	4	−2	4
3	1	2	4
4	5	−1	1
5	3	2	4
6	8	−2	4
7	7	0	0
8	6	2	4
			$\sum d_i^2 = 22$

Here, n = number of paired observations = 8

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 22}{8 \times 63} = 1 - \frac{132}{504} = \frac{372}{504} = 0.74$$

Sometimes we do not have rank but actual values of variables are available. If we are interested in rank correlation coefficient, we find ranks from the given values. Considering this case we are taking a problem and try to solve it.

Example 2: Suppose we have ranks of 5 students in three subjects Computer, Physics and Statistics and we want to test which two subjects have the same trend.

Rank in Computer	2	4	5	1	3
Rank in Physics	5	1	2	3	4
Rank in Statistics	2	3	5	4	1

Solution: In this problem, we want to see which two subjects have same trend i.e. which two subjects have the positive rank correlation coefficient.

Here we have to calculate three rank correlation coefficients

r_{12s} = Rank correlation coefficient between the ranks of Computer and Physics

r_{23s} = Rank correlation coefficient between the ranks of Physics and Statistics

r_{13s} = Rank correlation coefficient between the ranks of Computer and Statistics

Let R_1, R_2 and R_3 be the ranks of students in Computer, Physics and Statistics respectively.

Let R_1, R_2 and R_3 be the ranks of students in Computer, Physics and Statistics respectively.

Rank in Computer (R_1)	Rank in Physics (R_2)	Rank in Statistics (R_3)	$d_{12}=R_1-R_2$	d_{12}^2	$d_{23}=R_2-R_3$	d_{23}^2	$d_{13}=R_1-R_3$	d_{13}^2
2	5	2	−3	9	3	9	0	0
4	1	3	3	9	−2	4	1	1
5	2	5	3	9	−3	9	0	0
1	3	4	−2	4	−1	1	−3	9
3	4	1	−1	1	−3	9	2	4
Total				32		32		14

Thus,

$$\sum d_{12}^2 = 32,\ \sum d_{23}^2 = 32 \text{ and } \sum d_{13}^2 = 14.$$

Now

$$r_{12s} = 1 - \frac{6\sum d_{12}^2}{n(n^2-1)} = 1 - \frac{6 \times 32}{5 \times 24} = 1 - \frac{8}{5} = -\frac{3}{5} = -0.6$$

$$r_{23s} = 1 - \frac{6\sum d_{23}^2}{n(n^2-1)} = 1 - \frac{6 \times 32}{5 \times 24} = 1 - \frac{8}{5} = -\frac{3}{5} = -0.6$$

$$r_{13s} = 1 - \frac{6\sum d_{13}^2}{n(n^2-1)} = 1 - \frac{6 \times 14}{5 \times 24} = 1 - \frac{7}{10} = \frac{3}{10} = 0.3$$

r_{12s} is negative which indicates that Computer and Physics have opposite trend. Similarly, negative rank correlation coefficient r_{23s} shows the opposite

3. Calculate tied rank correlation coefficient from the following data

Expenditure on advertisement	10	15	14	25	14	14	20	22
Profit	6	25	12	18	25	40	10	7

$$r_s = 1 - \frac{6\left\{83.50 + \frac{3 \times 8}{12} + \frac{2 \times 3}{12}\right\}}{8X\,63}$$

$$r_s = 1 - \frac{6(83.50 + 2.50)}{504}$$

$$r_s = 1 - \frac{516}{504}$$

$$r_s = 1 - 1.024 = -0.024$$

There is a negative association between expenditure on advertisement and profit.

Regression Lines
Correlation describes the strength of an association between two variables and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A. However, if the two variables are related it means that when one changes by a certain amount the other changes on an average by a certain amount.
If y represents the dependent variable and x the independent variable, this relationship is described as the regression of y on x.

The relationship can be represented by a simple equation called the regression equation. In this context "regression" (the term is a historical anomaly) simply means that the average value of y is a "function" of x, that is, it changes with x.

The regression equation representing how much y changes with any given change of x can be used to construct a regression line on a scatter diagram, and in the simplest case this is assumed to be a straight line.

Properties of the Regression coefficient

- It is denoted by b
- Both regression coefficients must have the same sign. If b_{yx} is positive, b_{xy} will also be positive and vice versa.
- If one regression coefficient is greater than one then the other one is less than one.

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

Regression coefficients are classified as:
(1) Simple, partial and multiple
(2) Positive and negative and
(3) Linear and non-linear.

Regression Coefficients are

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad , \quad \text{where} \quad \sigma_x = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \qquad \sigma_y = \sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2}$$

and $$r = \frac{\frac{\sum xy}{n} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}$$

The regression line of X on Y is

$$x-\bar{x}=b_{xy}\left(y-\bar{y}\right)$$

The regression line of Y on X is

$$y-\bar{y}=b_{yx}\left(x-\bar{x}\right)$$

Problems. 1

Find the regression lines:

X	6	8	10	18	20	23
Y	40	36	20	14	10	2

Solution

X	Y	X^2	Y^2	XY
6	40	36	1600	240
8	36	64	1296	288
10	20	100	400	200
18	14	324	196	252
20	10	400	100	200
23	2	529	4	46
$\sum X=85$	$\sum Y=122$	$\sum X^2=1453$	$\sum Y^2=3596$	$\sum XY=1226$

$$\bar{X}=\frac{\sum x}{n}=\frac{85}{6}=14.17,\ \bar{Y}=\frac{\sum y}{n}=\frac{122}{6}=20.33$$

$$\sigma_x=\sqrt{\frac{\sum x^2}{n}-\left(\frac{\sum x}{n}\right)^2}=\sqrt{\frac{1453}{6}-\left(\frac{85}{6}\right)^2}=6.44$$

$$\sigma_y=\sqrt{\frac{\sum y^2}{n}-\left(\frac{\sum y}{n}\right)^2}=\sqrt{\frac{3596}{6}-\left(\frac{122}{6}\right)^2}=13.63$$

$$r=\frac{\frac{\sum xy}{n}-\bar{x}\bar{y}}{\sigma_x\sigma_y}=\frac{\frac{1226}{6}-(14.17)(20.33)}{(6.44)(13.63)}=-0.95$$

$$b_{xy}=r\frac{\sigma_x}{\sigma_y}=-0.95\times\frac{6.44}{13.63}=-0.45$$

$$b_{yx}=r\frac{\sigma_y}{\sigma_x}=-0.95\times\frac{13.63}{6.44}=-2.01$$

The regression line X on Y is

$$x-\bar{x}=b_{xy}\left(y-\bar{y}\right)\Rightarrow x-14.17=-0.45\left(y-\bar{y}\right)$$

$$\Rightarrow x=-0.45y+23.32$$

The regression line Y on X is

$$y-\bar{y}=b_{yx}\left(x-\bar{x}\right)\Rightarrow y-20.33=-2.01\left(x-14.17\right)$$

$$\Rightarrow y=-2.01x+48.81$$

2. Using the given information compute Also compute $\bar{x},\bar{y}$ and r. σ_X when $\sigma_Y=2$.

$2x+3y=8$ and $4x+y=10$.

Solution

$$2x+3y=8 \text{ ---------- (1)}$$

$$4x+y=10 \text{ ---------- (2)}$$

$$(1)\times 2\Rightarrow 4x+6y=16 \text{ ------- (3)}$$

$$(2)-(3)\Rightarrow -5y=-6$$

$$\Rightarrow y=\frac{6}{5}$$

Equation $(1)\Rightarrow 2x+3\left(\frac{6}{5}\right)=8$

$$\Rightarrow 2x=8-\frac{18}{5}$$

$$\Rightarrow x=\frac{11}{5}$$

i.e. $\bar{x}=\frac{11}{5}$ & $\bar{y}=\frac{6}{5}$

To find r, Let $2x+3y=8$ be the regression equation of X on Y.

$$2x=8-3y\Rightarrow x=4-\frac{3}{2}y$$

$\Rightarrow b_{xy}=$ Coefficient of Y in the equation of X on $Y=-\frac{3}{2}$

Let $4x+y=10$ be the regression equation of Y on X

$$\Rightarrow y=10-4x$$

$\Rightarrow b_{yx}=$ coefficient of X in the equation of Y on $X=-4$.

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

$$=-\sqrt{\left(-\frac{3}{2}\right)(-4)} \qquad \left(\because b_{xy} \text{ \& } b_{yx} \text{ are negative }\right)$$

$$=-2.45$$

Since r is not in the range of $(-1 \le r \le 1)$ the assumption is wrong.

Now let equation (1) be the equation of Y on X

$$\Rightarrow y = \frac{8}{3} - \frac{2x}{3}$$

$$\Rightarrow b_{yx} = \text{Coefficient of } X \text{ in the equation of } Y \text{ on } X$$

$$b_{yx} = -\frac{2}{3}$$

from equation (2) be the equation of X on Y

$$b_{xy} = -\frac{1}{4}$$

$$r = \pm\sqrt{b_{xy} b_{yx}} \quad = \sqrt{-\frac{2}{3} \times -\frac{1}{4}} = 0.4081$$

To compute σ_y from equation (4) $b_{yx} = -\frac{2}{3}$

But we know that $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

$$\Rightarrow -\frac{2}{3} = 0.4081 \times \frac{\sigma_y}{2}$$

$$\Rightarrow \sigma_y = -3.26$$

Partial Correlation Coefficient:

Partial correlation coefficient provides a measure of the relationship between the dependent variable and other variables, with the effect of the most of the variables eliminated.

Let $r_{12.3}$ be the coefficient of partial correlation between X_1 and X_2 keeping X_3 constant, then

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$$

Similarly,

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{23}^2\right)}}$$

where $r_{13.2}$ is the coefficient of partial correlation between X_1 and X_3 keeping X_2 constant.

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{\left(1 - r_{12}^{\,2}\right)\left(1 - r_{13}^{\,2}\right)}}$$

where $r_{23.1}$ is the coefficient of partial correlation between X_2 and X_3 keeping X_1 constant.

Problems:

1. If $r_{12} = 0.8$, $r_{13} = 0.4$ and $r_{23} = 0.56$, find the value of $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$

Solution:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1 - r_{13}^{\,2}\right)\left(1 - r_{23}^{\,2}\right)}}$$

Substituting the given values,

$$r_{12.3} = \frac{0.8 - 0.4 \times 0.56}{\sqrt{1-(0.4)^2}\sqrt{1-(0.56)^2}}$$

$$= 0.7586$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{\left(1 - r_{12}^{\,2}\right)\left(1 - r_{23}^{\,2}\right)}}$$

$$= \frac{0.4 - (0.8)(0.56)}{\sqrt{1-(0.8)^2}\sqrt{1-(0.56)^2}}$$

$$= -0.0966$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{\left(1 - r_{12}^{\,2}\right)\left(1 - r_{13}^{\,2}\right)}}$$

$$= \frac{0.56 - (0.8)(0.4)}{\sqrt{1-(0.8)^2}\sqrt{1-(0.4)^2}}$$

$$= 0.4364$$

2. The correlation between a general intelligence test and school achievement in a group of children from 6 to 15 years old is 0.80. The correlation between the general intelligence test and age in the same group is 0.70 and the correlation between school achievement and age is 0.60. What is the correlation between general intelligence and school achievement in children of the same age?

Solution:

Let X_1 denote general intelligence tes

X_2 denote school achievement.

X_3 denote age.

We are given $r_{12} = 0.8$, $r_{13} = 0.7$ and $r_{23} = 0.6$

We have to find $r_{12.3}$

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1 - r_{13}^{\ 2}\right)\left(1 - r_{23}^{\ 2}\right)}}$$

$$= \frac{0.8 - (0.7)(0.6)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.6)^2}}$$

$$= 0.6651$$

Multiple Correlation

The coefficient of multiple linear correlation is represented by R, and it is common to add subscripts designating the variable involved. Thus $R_{1.23}$ would represent the coefficient of multiple linear correlation between X_1, on the one hand, and X_2 and X_3 on the other hand. The subscript of the dependent variable is always to the left of the point.

The coefficient of multiplication correlation can be expressed in terms of r_{12}, r_{13} and r_{23} as follows:

$$R_{1.23} = \sqrt{\frac{r_{12}^{\ 2} + r_{13}^{\ 2} - 2r_{12}r_{13}r_{23}}{1 - r_{23}^{\ 2}}}$$

$$R_{2.13} = \sqrt{\frac{r_{12}^{\ 2} + r_{23}^{\ 2} - 2r_{12}r_{13}r_{23}}{1 - r_{13}^{\ 2}}}$$

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

Problems 1

If $r_{12} = 0.09$, $r_{13} = 0.75$ and $r_{23} = 0.7$, find $R_{1.23}$

Solution

$$R_{1.23} = \sqrt{\frac{r_{12}^{\,2} + r_{13}^{\,2} - 2r_{12}r_{13}r_{23}}{1 - r_{23}^{\,2}}}$$

Substituting the given values,

$$R_{1.23} = \sqrt{\frac{0.9^2 + 0.75^2 - 2 \times 0.9 \times 0.75 \times 0.7}{1 - 0.7^2}}$$

$$= 0.9156$$

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – IV– SMTA 1308 – Mathematical Statistics

UNIT IV STANDARD DISTRIBUTIONS

Binomial, Poisson, Normal, and Uniform distributions - Geometric, Exponential, Gamma and Beta distributions. The interrelationship between distributions.

Discrete Distributions

1. Binomial Distribution
2 Poisson Distribution
3. Geometric Distribution

Binomial Distribution

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

$$P(X=x)=p(x)=\begin{cases} nC_x p^x q^{n-x}, x=0,1,2,\ldots,n; q=1-p \\ 0, otherwise \end{cases}$$

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

Problem.1

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

Solution:

M.G.F.of Binomial distribution:-

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

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

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

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

Mean of Binomial distribution

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

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

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

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

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

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

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

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

Problem.2

Comment the following: "The mean of a binomial distribution is 3 and variance is 4

Solution:

In binomial distribution, mean > variance but Variance < Mean

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

Problem.3

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

Solution:

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

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

Problem.4

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

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

Solution:

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

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

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

Poisson distribution:

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

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

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

Poisson distribution as limiting form of binomial distribution:

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

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

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

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

Proof:

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

Let $np=\lambda$

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

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

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

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

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

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

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

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

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

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

Geometric distribution:

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

$$P\left(X=x\right)=\begin{cases}q^{x-1}p;x=1,2,\ldots;0<p\le 1\\0,otherwise\end{cases}$$

Problem.1

Criticise the following statement: "The mean of a Poisson distribution is 5 while the

Problem 2

Find the Moment generating function of geometric distribution and find its Mean and Variance

Solution:

t is

$$M_X(t) = E\left(e^{tX}\right)$$

$$= \sum_{x=1}^{\infty} e^{tx}\, q^{x-1} p$$

$$= \sum_{x=1}^{\infty} p e^t \left(q e^t\right)^{x-1}$$

$$= p e^t \left(1 + q e^t + \left(q e^t\right)^2 + \cdots\right)$$

$$= p e^t \left(1 - q e^t\right)^{-1}$$

$$M_X(t) = \frac{p e^t}{1 - q e^t}$$

$$\mu_1' = M_x'(0) = \left[\frac{d}{dt}\left(\frac{p e^t}{\left(1 - q e^t\right)}\right)\right]_{t=0} = \left[\left(\frac{p e^t}{\left(1 - q e^t\right)^2}\right)\right]_{t=0} = \frac{1}{p}$$

$$\mu_2' = M_x''(0) = \left[\frac{d^2}{dt^2}\left(\frac{p e^t}{\left(1 - q e^t\right)}\right)\right]_{t=0} = \left[\frac{d}{dt}\left(\frac{p e^t}{\left(1 - q e^t\right)^2}\right)\right]_{t=0} = \frac{1+q}{p^2}$$

$$\text{Mean} = \mu_1' = \frac{1}{p}$$

$$\text{Variance} = \mu_2' - \left(\mu_1'\right)^2 = \frac{1+q}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{q}{p^2}$$

$$\therefore P[X>t]=\sum_{x=t+1}^{\infty} q^{x-1}p = q^t p + q^{t+1}p + q^{t+2}p + ... = q^t p\left[1+q+q^2+q^3+....\right]$$

$$= q^t p(1-q)^{-1} = q^t p(p)^{-1} = q^t$$

and 't'

Hence $P[X>s+t]=q^{s+t}$ and $P[X>s]=q^s$

$$(1) \Rightarrow P\left[X>s+t/X>s\right]=\frac{P[X>s+t\cap X>s]}{P[X>s]}=\frac{q^{s+t}}{q^s}=q^t=P[X>t]$$

$$\Rightarrow P\left[X>s+t/X>s\right]=P(X>t)$$

Problem.3

If the probability is $\frac{1}{4}$ that a man will hit a target what is the chance that he will hit the target for the first time in the 7th trial?

Solution:

The required probability is

$$P[FFFFFFS]=P(F)P(F)P(F)P(F)P(F)P(F)P(S)$$

$$=q^6 p=\left(\frac{3}{4}\right)^6.\left(\frac{1}{4}\right)=0.0445.$$

Problem.4

A die is cast until 6 appears what is the probability that it must cast more then five times?

Solution:

Probability of getting six $=\frac{1}{6}$

$$\therefore p=\frac{1}{6} \;\&\; q=1-\frac{1}{6}$$

$$=1-\sum_{x=1}^{5}\left(\frac{5}{6}\right)^{x-1}.\frac{1}{6}$$

$$=1-\left[\left(\frac{1}{6}\right)+\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)+\left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right)+\left(\frac{5}{6}\right)^3\left(\frac{1}{6}\right)+\left(\frac{5}{6}\right)^4\left(\frac{1}{6}\right)\right]$$

tric distribution

$$=1-\frac{\frac{1}{6}\left[1-\left(\frac{5}{6}\right)^5\right]}{1-\frac{5}{6}}=\left(\frac{5}{6}\right)^5=0.4019$$

Problem.5

Suppose that a trainee soldier shoots a target an independent fashion. If the probability that the target is shot on any one shot is 0.8.

(i) What is the probability that the target would be hit on 6th attempt?

(ii) What is the probability that it takes him less than 5 shots?

Solution:

Here $p = 0.8, q = 1 - p = 0.2$

$$P[X = x] = q^{x-1}p, x = 1, 2...$$

(i) The probability that the target would be hit on the 6th attempt $= P[X = 6]$

$$= (0.2)^5 (0.8) = 0.00026$$

(ii) The probability that it takes him less than 5 shots $= P[X < 5]$

$$= \sum_{x=1}^{4} q^{x-1}p = 0.8\sum_{x=1}^{4}(0.2)^{x-1}$$

$$= 0.8[1 + 0.2 + 0.04 + 0.008] = 0.9984$$

Continuous Distributions

1. Uniform Distribution
2. Exponential Distribution
3. Uniform Distribution
4. Normal Distribution
5. Gamma Distribution
6. Beta Distribution

Uniform (or) Rectangular distribution:

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

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

Problem.1

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

Solution:

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

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

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

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

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

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

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

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

Exponential distribution:

A continuous random variable X assuming non negative values is said to have an exponential distribution with parameter $\theta>0$, if its probability density function is given by

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

Problem 1

Find the moment generating function of Exponential distribution and find its mean and variance.

Solution:

We know that $f(x)=\begin{cases}\lambda e^{-\lambda x}, x\geq 0\\ 0, otherwise\end{cases}$

$$M_X(t)=E(e^{tx})=\int_0^{\infty}e^{tx}f(x)dx=\int_0^{\infty}\lambda e^{-\lambda x}e^{tx}dx$$

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

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

$$\text{Mean} = \mu_1' = \left[\frac{d}{dt} M_X(t) \right]_{t=0} = \left[\frac{\lambda}{(\lambda - t)^2} \right]_{t=0} = \frac{1}{\lambda}$$

$$\mu_2' = \left[\frac{d^2}{dt^2} M_X(t) \right]_{t=0} = \left[\frac{\lambda(2)}{(\lambda - t)^3} \right]_{t=0} = \frac{2}{\lambda^2}$$

$$\text{Variance} = \mu_2' - (\mu_1')^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Problem.2

State and prove the memoryless property of exponential distribution.

Solution:

Statement:

If X is exponentially distributed with parameters λ, then for any two positive integers's' and't', $P[x > s+t / x > s] = P[x > t]$

Proof:

Let X denote the time to failure of the component then X has exponential distribution with $Mean = 1000$ hours.

$$\therefore \frac{1}{\lambda} = 10,000 \Rightarrow \lambda = \frac{1}{10,000}$$

The p.d.f. of X is $f(x) = \begin{cases} \frac{1}{10,000} e^{-\frac{x}{10,000}}, & x \geq 0 \\ 0, & otherwise \end{cases}$

$$P[x > s]$$

$$= \frac{P[X > s+t]}{P[X > s]} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}$$

$$= P[x > t]$$

Problem.3

A component has an exponential time to failure distribution with mean of 10,000 hours.

(i). The component has already been in operation for its mean life. What is the probability that it will fail by 15,000 hours?

(ii). At 15,000 hours the component is still in operation. What is the probability that it will operate for another 5000 hours.

Solution:

(i) Probability that the component will fail by 15,000 hours given it has already been in operation for its mean life $= P[x < 15,000 / x > 10,000]$

$$= \frac{P[10,000 < X < 15,000]}{P[X > 10,000]}$$

$$= \frac{\int_{10,000}^{15,000} f(x)\,dx}{\int_{10,000}^{\infty} f(x)\,dx} = \frac{e^{-1} - e^{-1.5}}{e^{-1}}$$

$$= \frac{0.3679 - 0.2231}{0.3679} = 0.3936.$$

(ii) Probability that the component will operate for another 5000 hours given that it is in operational 15,000 hours $= P[X > 20,000 / X > 15,000]$

$$= P[x > 5000] \quad \text{[By memoryless prop]}$$

$$= \int_{5000}^{\infty} f(x)\,dx = e^{-0.5} = 0.6065$$

Normal distribution:

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

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

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

Problem.1

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

Solution:

Moment generating function of Normal distribution

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

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

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

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

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

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

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

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

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

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

Problem.2

State and prove the additive property of normal distribution.

Statement

If $X_1, X_2, ..., X_n$ are n independent normal random variates with mean (μ_1, σ_1^2), $(\mu_2, \sigma_2^2), ... (\mu_n, \sigma_n^2)$ then $X_1 + X_2 + ... + X_n$ also a normal random variable with mean $\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$.

Proof

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

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

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

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

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

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

This proves the property.

Problem.3

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

Solution:

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

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

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

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

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

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

$$= 0.2881 + 0.4772 = 0.7653.$$

Problem.4

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

Solution:

Let X be marks of the candidates

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

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

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

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

Problem.5

Given that X is normally distribution with mean 10 and probability $P[X > 12] = 0.1587$.

What is the probability that X will fall in the interval $(9,11)$.

Solution:

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

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

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

$$= 50 - 1.9 = 48.1$$

The average mark should be 48 nearly.

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

Then $P[X > 12] = 0.1587$

$P[Z > Z_1] = 0.1587$

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

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

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

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

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

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

$$\begin{aligned} \therefore P[9 < X < 11] &= P[-0.5 < z < 0.5] \\ &= 2P[0 < z < 0.5] \\ &= 2 \times 0.1915 = 0.3830 \end{aligned}$$

Problem 6

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

Solution:

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

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

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

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

$\therefore z_1 = 0.495$

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

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

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

$\therefore z_2 = 1.405$

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

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

Gamma or Erlang Distribution

A continuous RV X is said to follow a general gamma distribution with parameters λ>0 and k>0 if the probability density function

$$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\overline{|k}} \; for x \le 0$$

If λ=1, then Erlang distribution is gamma distribution.

MGF of Gamma Distribution

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-k}$$

Mean and Variance

$$E(X) = \frac{k}{\lambda}$$

Variance (x)= $E(X^2)$-$[E(X)]^2$

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

Beta Distribution

A continuous RV X is said to follow Beta distribution if it has the pdf

$$f(x) = \frac{(x-a)^{p-1}(b-x)^{q-1}}{B(p,q)(b-a)^{p+q-1}} \qquad a \le x \le b; p, q > 0$$

where p and q are the parameters and a and b are lower and upper bound of the distribution. B(p,q) is the beta function.

The case where $a = 0$ and $b = 1$ is called the **standard beta distribution**. The equation for the standard beta distribution is

$$f(x) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)} \qquad 0 \le x \le 1; p, q > 0$$

CDF

The formula for the cumulative distribution function of the beta distribution is also called the incomplete beta function ratio (commonly denoted by I_x) and is defined as

$$F(x) = I_x(p, q) = \frac{\int_0^x t^{p-1}(1-t)^{q-1}dt}{B(p,q)} \quad 0 \le x \le 1; p, q > 0$$

where B is the beta function defined above.

The formulas below are for the case where the lower limit is zero and the upper limit is one.

Mean	$\frac{p}{p+q}$
Mode	$\frac{p-1}{p+q-2}$ $\quad p, q > 1$
Range	0 to 1
Standard Deviation	$\sqrt{\frac{pq}{(P+q)^2(p+q+1)}}$
Coefficient of Variation	$\sqrt{\frac{q}{p(p+q+1)}}$
Skewness	$\frac{2(q-p)\sqrt{p+q+1}}{(p+q+2)\sqrt{pq}}$

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – V – SMTA1308 – Mathematical Statistics

UNIT V SAMPLING THEORY

Introduction-large sample test based on normal distribution - Sampling distribution of the mean – Confidence limits Test for the single mean difference between means, proportion, the difference between proportion - small sample test based on t, F distributions - Test for a single mean, the difference between means, standard deviation, the difference between standard deviations – Chi-square test for goodness of fit, independence of attributes.

Population: The group of individuals, under study is called population.

Sample: A finite subset of statistical individuals in a population is called Sample.

Sample size: The number of individuals in a sample is called the Sample size.

Parameters and Statistics: The statistical constants of the population are referred as Parameters and the statistical constants of the Sample are referred to as Statistics.

Standard Error: The standard deviation of sampling distribution of a statistic is known as its standard error and is denoted by (S.E)

Test of Significance: It enables us to decide on the basis of the sample results if the deviation between the observed sample statistic and the hypothetical parameter value is significant or the deviation between two sample statistics is significant.

Null Hypothesis: A definite statement about the population parameter which is usually a hypothesis of no-difference and is denoted by Ho.

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

if $\mu = \mu 0$ is the null hypothesis H0 then, the alternate hypothesis H1 could be $\mu > \mu 0$ (Right tail) or $\mu < \mu 0$ (Left tail) or $\mu \neq \mu 0$ (Two tail test)

Errors in Sampling: Type I and Type II errors.

Type I error: Rejection of H0, when it is true.

Type II error: Acceptance of H0, when it is false.

Critical region: A region corresponding to a statistic “t” in the sample space S which leads to the rejection of Ho is called Critical region or Rejection region.

Acceptance Region: Those regions which lead to the acceptance of Ho are called Acceptance Region.

Level of Significance: The probability α that a random value of the statistic “t” belongs to the critical region is known as the level of significance.

Types of samples: Small sample and Large sample. A sample is said to be a small sample if the size is less than or equal to 30 otherwise it is a large sample.

Large Sample

Z test for mean

Test of significance for single Mean

$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$, where $\bar{x}$ the sample mean, μ is is the population mean, σ is the population standard deviation and n is the sample size.

Test of significance for difference of mean

$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$, where $\bar{x}_1$ is the first sample mean, $\bar{x}_2$ is the second sample mean, n_1 is the first sample size, n_2 is the second sample size, s_1^2 is the first sample variance and s_2^2 is the second sample variance.

Confidence Limits

The values of $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ are called 95% confidence limits for the mean of the population corresponding to the given sample. The values of $\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$ are called 99% confidence limits for the mean of the population corresponding to the given sample.

Z test for proportions

Test of significance for single proportion

$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$, where P is the population proportion, Q = 1 – P, p is the sample proportion and n is the sample size.

Test of significance for difference of proportion

$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, where p_1 is the first sample proportion, p_2 is the second sample proportion, n_1 is the first sample size, n_2 is the second sample size, $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

and Q = 1 – P

Small Sample

t -Test of significance for single Mean

$t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n-1}}}$, where $\bar{x}$ the sample mean, μ is is the population mean, s is the sample standard deviation and n is the sample size.

If the mean and standard deviation are not given, then the following formulae are used to calculate

$\bar{x} = \dfrac{\sum x}{n}$, $s^2 = \dfrac{\sum (x - \bar{x})^2}{n}$

Degrees of freedom is n - 1

Confidence Limits

Let $\bar{x}$ be the sample mean and n be the sample size. Let s be the sample standard deviation. Then the 95 % level confidence limits are given by $\bar{x} \pm t_{0.05} \dfrac{s}{\sqrt{n-1}}$. The 99 % level confidence limits are given by $\bar{x} \pm t_{0.01} \dfrac{s}{\sqrt{n-1}}$.

Test of significance for difference of mean

$Z = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, where $\bar{x}_1$ is the first sample mean, $\bar{x}_2$ is the second sample mean, n_1 is the first sample size, n_2 is the second sample size, $s^2 = \dfrac{n_1 {s_1}^2 + n_2 {s_2}^2}{n_1 + n_2 - 2}$.

Degrees of freedom is $n_1 + n_2 - 2$

F test

F= $\dfrac{Greater\ \text{var}iance}{Smaller\ \text{var}iance}$ i.e, $F = \dfrac{{S_1}^2}{{S_2}^2}$ if ${S_1}^2 \rangle {S_2}^2$ (OR) $F = \dfrac{{S_2}^2}{{S_1}^2}$ if ${S_2}^2 \rangle {S_1}^2$

If the sample variances s_1^2 and s_2^2 are given, then the following formula can be used to calculate S_1^2 and S_2^2 :

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1}, S_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$$

If the sample variances s_1^2 and s_2^2 are not given and the set of observations for both samples are given then the following formula can be used to calculate S_1^2 and S_2^2

$S_1^2 = \frac{\sum (x - \bar{x})^2}{n_1 - 1}$, $S_2^2 = \frac{\sum (y - \bar{y})^2}{n_2 - 1}$, where n_1 is the first sample size, n_2 is the second sample size, $\bar{x}$ is the first sample mean and $\bar{y}$ is the second sample mean.

χ^2 test

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where O is the observed frequency and E is the expected frequency.

Calculation of expected frequencies in testing independence of attributes

Expected Frequency = (Row total * Column Total) / Grand total

An explanation for the above with two classes is given below

Observed Frequencies

			Total
	a	c	a+c
	b	d	b+d
Total	a+b	c+d	a+b+c+d = N

Expected Frequencies

			Total
	E(a) = $\frac{(a+c)(a+b)}{N}$	E(c) = $\frac{(a+c)(c+d)}{N}$	a+c
	E(b) = $\frac{(b+d)(a+b)}{N}$	E(d) = $\frac{(c+d)(b+d)}{N}$	b+d
Total	a+b	c+d	a+b+c+d = N

Problems

1. **A company manufacturing electric light bulbs claims that the average life of its bulbs is 1600 hours. The average life and standard deviation of a random sample of 100 such bulbs were 1570 hours and 120 hours respectively. Test the claim of the company at 5% level of significance.**

   Solution:

   Null Hypothesis H_0 : $\mu = 1600$. There is no significant difference between sample mean and population mean

   Alternative Hypothesis H_1 : $\mu \neq 1600$. There is a significant difference between sample mean and population mean.

   The statistic test is $Z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

   $$Z = \frac{1570 - 1600}{\frac{120}{\sqrt{100}}} = -2.5$$

   $$|z| = 2.5$$

   Calculated value z = 2.5

   Tabulated value of z at 5% level of significance for a two tail test is 1.96

   Calculated value >Tabulated value, H_0 is rejected.

   We cannot accept the claim of the company.

2. **The breaking strength of ropes produced by a manufacturer has mean 1800N and standard deviation 100N. By introducing a new technique in the manufacturing process it is claimed that the breaking strength has increased. To test this claim a sample of 50 ropes is tested and it is found that the breaking strength is 1850N. Can we support the claim at 1% level of significance?**

   Solution:

   Null Hypothesis H_0: $\mu = 1800$ N

   Alternative Hypothesis H_1: $\mu > 1800$ N (one tailed test)

   n = 50, $\bar{x} = 1850$ $\mu = 1800$ $\sigma = 100$

   The statistic test is $Z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

   $$Z = \frac{1850 - 1800}{\frac{100}{\sqrt{50}}} = 3.54$$

Calculated value z = 3.54
Tabulated value of z at 5% level of significance for a one tail test is 2.33
Calculated value >Tabulated value, H_0 is rejected.
The difference is significant and so we support the claim of the manufacturer.

3. **Measurements of the weights of a random sample of 200 ball bearings made by a certain machine during one week showed a mean of 0.824N and a standard deviation of 0.042N. Find the 95% and 99% confidence limits for the mean weight of all the ball bearings.**

Solution:

The 95% confidence limits are $\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$

n = 200, $\bar{x} = 0.824$ s = 0.042

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}} = 0.824 \pm (1.96)\left(\frac{0.042}{\sqrt{200}}\right) = 0.824 \pm 0.0058$$

The 95% confidence interval is (0.8182, 0.8298)

The 99% confidence limits are $\bar{x} \pm 2.58 \frac{s}{\sqrt{n}}$

$$\bar{x} \pm 2.58\frac{s}{\sqrt{n}} = 0.824 \pm (2.58)\left(\frac{0.042}{\sqrt{200}}\right) = 0.824 \pm 0.0077$$

The 99% confidence interval is (0.8163, 0.8317)

4. **In a survey of buying habits, 400 women shoppers are chosen at random in supermarket A. Their average weekly food expenditure is Rs.250 with standard deviation Rs.40. For 400 women shoppers chosen at random in supermarket B, the average weekly food expenditure is Rs.220 with standard deviation is Rs.55. Test at 1% level of significance whether the average weekly food expenditure of the populations of shoppers are equal.**

Solution:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

$n_1 = 400$ $n_2 = 400$ $\bar{x} = 250$ $\bar{y} = 220$ $s_1 = 40$ $s_2 = 55$

$$s = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)} = \sqrt{\frac{40^2}{400} + \frac{55^2}{400}} = 3.4$$

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{250 - 220}{3.4} = 8.82$$

Calculated value z = 8.82
Tabulated value of z at 1% level of significance for a two tailed test is 2.56

Calculated value >Tabulated value, H_0 is rejected.
The difference in the weekly food expenditure is significantly different.

5. **A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Test whether the proportion of bad ones is not significantly different from 0.1 at 1% level of significance**

Solution:

Null Hypothesis H_0: P = 0.1There is no significant difference between sample and population proportion.

Alternative Hypothesis H_1: $P \neq 0.1$ There is a significant difference between sample and population proportion.

The statistic test is $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}}$

$$p = \frac{65}{500} = 0.13$$

P = 0.1 Q= 1 – P = 1 – 0.1= 0.9

$$Z = \frac{0.13 - 0.1}{\sqrt{\dfrac{(0.1)(0.9)}{500}}} = 2.238$$

Calculated value z = 2.238

Tabulated value of z at 5% level of significance for a two tail test is 1.96

Calculated value >Tabulated value, H_0 is rejected.

The proportion of bad ones is significantly different from 0.1

6. **In a sample of 1000 people, 540 were rice eaters and the rest were wheat eaters. Can we assume that the proportion of rice eaters is more than 50% at 1% level of significance.**

Solution:

H_0: P = 0.5

H_1: P > 0.5 (One tailed test)

P = 0.5 Q= 1 – P = 1 – 0.5 = 0.5

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

The statistic test is $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}}$

$$Z = \frac{0.54 - 0.5}{\sqrt{\dfrac{(0.5)(0.5)}{1000}}} = 2.532$$

Calculated value z = 2.532

Tabulated value of z at 5% level of significance for a one tail test is 2.33

Calculated value >Tabulated value, H_0 is rejected.

The rice eaters are more than 50% of the population.

7. In a random sample of 900 votes, 55% are favored the Democratic candidate for the post of President. Test the hypothesis that the Democratic candidate has more chances of winning the President post.

Solution:

H_0: P = 0.5

H_1: P > 0.5 (Right tailed test)

P = 0.5 , Q= 1 – P = 1 – 0.5 = 0.5

$$p = \frac{55}{100} = 0.55$$

The statistic test is $Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$

$$Z = \frac{0.55 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{900}}} = 3$$

Calculated value z = 3

The tabulated value of z at 5% level of significance for a one-tail test is 2.33

Calculated value >Tabulated value, H_0 is rejected.

The Democratic candidate is having more chances to win the President's Post.

8. In a random sample of 1000 persons from town A, 400 are found to be consumers of wheat. In a sample of 800 from town B, 400 are found to be consumers of wheat. Do these data reveal a significant difference between town A and town B so far as the proportion of wheat consumers is concerned?

Solution:

H_0: Two towns do not differ much as far as the proportion of wheat consumption. $P_1 = P_2$

H_1: $P_1 \neq P_2$

The Statistic test is $Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$$p_1 = \frac{400}{1000} = 0.4 \qquad p_2 = \frac{400}{800} = 0.5$$

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{1000(0.4) + 800(0.5)}{1000 + 800} = 0.444$$

Q= 1 – P = 1 – 0.444 = 0.556

$$Z = \frac{0.4 - 0.5}{\sqrt{(0.444)(0.556)\left(\frac{1}{1000} + \frac{1}{800}\right)}} = \frac{0.1}{0.024} = 4.17$$

Calculated value z = 4.17

Tabulated value of z at 5% level of significance for a two tail test is 1.96

Calculated value >Tabulated value, H_0 is rejected.

Hence the data reveal a significant difference between town A and town B so far as the proportion of wheat consumers is concerned.

9. In the past, a machine has produced washers having a thickness of 0.050 inch. To determine whether the machine is in proper working order, a sample of 10 washers is chosen, for which the mean thickness is 0.053 inch and the standard deviation is 0.003 inch. Test the hypothesis that the machine is in proper working order, using 5% and 1% level of significance.

Solution:

H_0: $\mu = 0.050$

H_1: $\mu \neq 0.050$ (two tailed test)

$n = 10 \qquad \bar{x} = 0.053 \qquad s = 0.003 \qquad \mu = 0.050$

The statistic test is $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n-1}}} = \dfrac{0.053 - 0.050}{\dfrac{0.003}{\sqrt{10-1}}} = 3.00$

Calculated value t = 3.00

Degree of freedom = n – 1 =10 – 1 = 9

At 5% LOS:

Tabulated value of t at 5% level of significance with 9 degrees of freedom for a two tailed test is 2.26

Calculated value >Tabulated value, H_0 is rejected.

The Machine is not in proper working order at 5% level of significance

Tabulated value of t at 1% level of significance with 9 degrees of freedom for a two tailed test is 3.25

Calculated value < Tabulated value, H_0 is accepted.

The Machine is in proper working order at 1% level of significance.

10. The specifications for a certain kind of ribbon call for a mean breaking strength of 180 pounds. If five pieces of the ribbon (randomly selected from the different rolls) have a mean breaking strength of 169.5 pounds with a standard deviation of 5.7 pounds. Test the null hypothesis $\mu = 180$ pounds against the alternative hypothesis $\mu < 180$ pounds at the 0.01 level of significance. Assume that the population distribution is normal.

Solution:

H_0: $\mu = 180$

H_1: $\mu < 180$ (left tailed test)

$n = 5 \qquad \bar{x} = 169.5 \qquad s = 5.7 \qquad \mu = 180$

The statistic test is $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n-1}}} = \dfrac{169.5 - 180}{\dfrac{5.7}{\sqrt{5-1}}} = -3.68$

Calculated value t = 3.68

Degree of freedom = n – 1 =5 – 1 = 4

Tabulated value of t at 1% level of significance with 4 degrees of freedom for a one tail test is 3.747.
Calculated value > Tabulated value, H_0 is accepted.
Hence the mean breaking strength can be taken as 180 pounds.

11. Ten individuals are chosen at random from a normal population and their heights are found to be 63,63,66,67,68,69,70,70,71,71 inches. Test the hypothesis that the mean height is greater than 66 inches at 5% level of significance

Solution:

H_0: $\mu = 66$

H_1: $\mu > 66$ (one tailed test)

$$\bar{x} = \frac{\sum x}{n} = \frac{678}{10} = 67.8$$

X	63	63	66	67	68	69	70	70	71	71	Total
$(x-\bar{x})$	- 4.8	- 4.8	- 1.8	- 0.8	0.2	1.2	2.2	2.2	3.2	3.2	
$(x-\bar{x})^2$	23.04	23.04	3.24	0.64	0.04	1.44	4.84	4.84	10.24	10.24	81.6

$$s^2 = \frac{\sum (x-\bar{x})^2}{n} = \frac{81.6}{10} = 8.16$$

s = 2.857

The statistic test is $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}} = \frac{67.8-66}{\frac{2.857}{\sqrt{9}}} = 1.89$

Calculated value t = 1.89
Degree of freedom = n – 1 =10 – 1 = 9
Tabulated value of t at 5% level of significance with 9 degrees of freedom for a one tail test is 1.833
Calculated value >Tabulated value, H_0 is rejected. Accepted H_1
The Mean is significantly higher than 66 inches.

12. Two independent samples of size 8 and 7 items had the following values

Sample I	**9**	**11**	**13**	**11**	**15**	**9**	**12**	**14**
Sample II	**10**	**12**	**10**	**14**	**9**	**8**	**10**	

Test if the difference between the mean is significant

Solution:

H_0: $\mu_1 = \mu_2$ There is no significant difference between means

H_1: $\mu_1 \neq \mu_2$ There is a significant difference between means

$$\bar{x} = \frac{\sum x}{n} = \frac{94}{8} = 11.75 \qquad \bar{y} = \frac{\sum y}{n} = \frac{73}{7} = 10.43$$

x	$(x-\bar{x})$	$(x-\bar{x})^2$	y	$(y-\bar{y})$	$(y-\bar{y})^2$
9	- 2.75	7.56	10	- 0.43	0.185
11	- 0.75	0.56	12	1.57	2.465
13	1.25	1.56	10	- 0.43	0.185
11	- 0.75	0.56	14	3.47	12.041
15	3.25	10.56	9	- 1.43	2.045
9	- 2.75	7.56	8	- 2.43	5.905
12	0.25	0.06	10	- 0.43	0.185
14	2.25	5.06			
94		33.48	73		23.011

$$s^2 = \frac{\sum (x-\bar{x})^2 + \sum (y-\bar{y})^2}{n_1 + n_2 - 2} = \frac{33.48 + 23.011}{8 + 7 - 2} = 4.35$$

s = 2.086

The statistic test is $t = \dfrac{\bar{x} - \bar{y}}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{11.75 - 10.43}{2.086\sqrt{\dfrac{1}{8} + \dfrac{1}{7}}} = 1.22$

Calculated value t = 1.22

Degree of freedom = $n_1 + n_2 - 2 = 8 + 7 - 2 = 13$

Tabulated value of t at 5% level of significance with 13 degrees of freedom for a two tail tets is 2.16

Calculated value < Tabulated value, H_0 is accepted

There is no significant difference between means.

13. The IQ of 16 students from one area of a city showed a mean of 107 with the standard deviation 10, while the IQ of 14 students from another area showed a mean of 112 with standard deviation 8. Is there a significant difference between the IQ's of the two groups at 1% and 5% level of significance?

Solution:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

$n_1 = 16 \quad n_2 = 14 \quad s_1 = 10 \quad s_2 = 8 \quad \bar{x} = 107 \quad \bar{y} = 112$

$$s^2 = \frac{n_1 s_1^{\,2} + n_2 s_2^{\,2}}{n_1 + n_2 - 2} = \frac{16(10)^2 + 14(8)^2}{16 + 14 - 2} = \frac{2496}{28} = 89.143$$

s = 9.44

The statistic test

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{107 - 112}{9.44\sqrt{\frac{1}{16} + \frac{1}{14}}} = -1.45$$

Calculated value t = 1.45

Degree of freedom = $n_1 + n_2 - 2 = 16 + 14 - 2 = 28$

At 5% LOS:

Tabulated value of t at 5% level of significance with 28 degree of freedom for a two tail test is 2.05

Calculated value < Tabulated value, H_0 is accepted

There is no significant difference in the IQ level of the two groups.

At 1% LOS:

Tabulated value of t at 1% level of significance with 28 degree of freedom is 2.76

Calculated value < Tabulated value, H_0 is accepted.

There is no significant difference in the IQ level of the two groups.

14. A random sample of 10 parts from machine A has a sample standard deviation of 0.014 and another sample of 15 parts from machine B has a sample standard deviation of 0.08. Test the hypothesis that the samples are from a population with same variance.

Solution:

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

$n_1 = 10 \quad n_2 = 15 \quad s_1 = 0.014 \quad s_2 = 0.08$

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{10 \times (0.014)^2}{10 - 1} = 0.0002$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{15 \times (0.08)^2}{15 - 1} = 0.006$$

$$F = \frac{S_2^2}{S_1^2} = \frac{0.006}{0.0002} = 30$$

Calculated value F = 30

Tabulated Value of F at 5% level of significant with (14, 9) degrees of freedom is 3.03

Calculated value > Tabulated value, H_0 is rejected

There is a significant difference in the variances of two populations.

15. Two random samples drawn from two normal populations are

Sample I	**20**	**16**	**26**	**27**	**23**	**22**	**18**	**24**	**25**	**19**		

Sample II	**27**	**33**	**42**	**35**	**32**	**34**	**38**	**28**	**41**	**43**	**30**	**37**

Obtain the estimates of the variances of the population and test whether the two populations have the same variance.

Solution:

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

$$\bar{x} = \frac{\sum x}{n} = \frac{220}{10} = 22$$

$$\bar{y} = \frac{\sum y}{n} = \frac{420}{12} = 35$$

x	(x-$\bar{x}$)	$(x-\bar{x})^2$	y	(y-$\bar{y}$)	$(y-\bar{y})^2$
20	- 2	4	27	- 8	64
16	- 6	36	33	- 2	4
26	4	16	42	7	49
27	5	25	35	0	0
23	1	1	32	- 3	9
22	0	0	34	- 1	1
18	- 4	16	38	3	9
24	2	4	28	- 7	49
25	3	9	41	6	36
19	- 3	9	43	8	64
			30	- 5	25
			37	2	4
		120			314

$$n_1 = 10 \qquad n_2 = 12$$

$$S_1^{\,2} = \frac{\sum (x-\bar{x})^2}{n_1 - 1} = \frac{120}{9} = 13.33$$

$$S_2^{\,2} = \frac{\sum (y-\bar{y})^2}{n_2 - 1} = \frac{314}{11} = 28.54$$

$$F = \frac{S_2^2}{S_1^2} = \frac{28.54}{13.33} = 2.14$$

Calculated value F = 2.14

Tabulated Value of F at 5% level of significance with (11, 9) degrees of freedom is 3.1

Calculated value < Tabulated value, H_0 is accepted
There is no significant difference between variances.

16. In one sample of 8 observations the sum of squares of deviations of the sample values from the sample mean was 84.4 and in another sample of 10 observations it was 102.6. Test whether this difference is significant at 5% level.

Solution:

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

$n_1 = 8 \quad n_2 = 10 \quad \sum (x - \bar{x})^2 = 84.4 \quad \sum (y - \bar{y})^2 = 102.6$

$$S_1^{\,2} = \frac{\sum (x - \bar{x})^2}{n_1 - 1} = \frac{84.4}{7} = 12.057$$

$$S_2^{\,2} = \frac{\sum (y - \bar{y})^2}{n_2 - 1} = \frac{102.6}{9} = 11.4$$

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

Calculated value F = 1.057
Tabulated Value of F at 5% level of significance with (7, 9) degrees of freedom is 3.29
Calculated value < Tabulated value, H_0 is accepted
There is no significant difference between variances.

17. The mean life of a sample of 9 bulbs was observed to be 1309 hrs with standard deviation 420 hrs. A second sample of 16 bulbs chosen from a different batch showed a mean life of 1205 hrs with a standard deviation 390 hrs. Test at 5% level whether both the samples come from the same normal population.

Solution:
Both t-test and F-test has to be done to check whether they have come from the same population. Fist F-test is done and the followed by t-test.
F-test:

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

$n_1 = 9 \quad n_2 = 16 \quad s_1 = 420 \quad s_2 = 390 \quad \bar{x} = 1309 \quad \bar{y} = 1205$

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{9 \times (420)^2}{9 - 1} = 198450$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{16 \times (390)^2}{16 - 1} = 162240$$

$$F = \frac{S_1^2}{S_2^2} = \frac{198450}{162240} = 1.223$$

Calculated value F = 1.223
Tabulated Value of F at 5% level of significant with (15, 8) degree of freedom is 3.22
Calculated value < Tabulated value, H_0 is accepted .

t-test:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

$$s^2 = \frac{n_1 {s_1}^2 + n_2 {s_2}^2}{n_1 + n_2 - 2} = \frac{9(420)^2 + 16(390)^2}{9 + 16 - 2} = \frac{4021200}{23} = 174834.7826$$

s = 418.13
The statistic test

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{1309 - 1205}{418.13\sqrt{\frac{1}{9} + \frac{1}{16}}} = \frac{104}{174.22} = 0.596$$

Calculated value t = 0.596
Degree of freedom = $n_1 + n_2 - 2$ = 9 + 16 – 2 = 23
Tabulated value of t at 5% level of significance with 23 degree of freedom is 2.069
Calculated value < Tabulated value, H_0 is accepted
Since in both F-Test and t-Test we have accepted the null hypothesis, we conclude that the samples have come from the same normal populations.

18. A dice is tossed 120 times with the following results:

No. turned up	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Frequency	**30**	**25**	**18**	**10**	**22**	**15**	**120**

Test the hypothesis that the dice is unbiased.

Solution:
Null Hypothesis H_0: The dice is an unbiased one.
Alternative Hypothesis H_1: The dice is biased

O	E	O – E	$(O-E)^2$	$\left[\frac{(O-E)^2}{E}\right]$
30	20	10	100	5.00
25	20	5	25	1.25
18	20	- 2	4	0.20
10	20	-10	100	5.00
22	20	2	4	0.20
15	20	- 5	25	1.25
				12.90

Calculated $\chi^2 = \left[\frac{(O-E)^2}{E}\right] = 12.90$

Degree of freedom = n – 1 = 6 – 1 = 5

Calculated value of χ^2 at 5% level of significance with 5 degree of freedom is 11.07

Tabulated value = 11.07

Tabulated value > calculated value, H_0 is rejected.

The dice are biased.

19. Genetic theory states that children having one parent of blood type M and other of blood type N will always be one of the three types M, MN, N and that the ratios of these types will be 1:2:1. A report states that out of 300 children having one M parent and one N parent, 30% were found to be of type M, 45% of type MN and remainder type N. Test the hypothesis using χ^2 test.

Solution:

H_0: There is no significant difference between the theoretical ratio and observed ratio.

H_1: There is no significant difference between the theoretical ratio and observed ratio.

If theoretical ratio is true the 300 children should be distributed as follows:

$$Type\ \ M = \frac{1}{4} \times 300 = 75$$

$$Type\ \ MN = \frac{2}{4} \times 300 = 150$$

$$Type\ \ N = \frac{1}{4} \times 300 = 75$$

Observed

$$Type\ \ M = \frac{30}{100} \times 300 = 90$$

$$Type\ \ MN = \frac{45}{100} \times 300 = 135$$

$$Type\ \ N = \frac{25}{100} \times 300 = 75$$

Type	observed	Expected	$O-E$	$(O-E)^2$	$\left[\frac{(O-E)^2}{E}\right]$
M	90	75	15	225	3
MN	135	150	- 15	225	1.5
N	75	75	0	0	0
Total					4.5

Calculated $\chi^2 = \left[\frac{(O-E)^2}{E}\right] = 4.5$

Degree of freedom = n – 1 = 3 – 1 = 2

Calculated value of χ^2 at 5% level of significance with 2 degree of freedom is 5.99

Tabulated value = 5.99

Calculated value < Tabulated value, H_0 is accepted

There is no significant difference between the theoretical ratio and observed ratio.

20. A certain drug was administered to 456 males, out of a total 720 in a certain locality, to test its efficacy against typhoid. To incidence of typhoid is shown below. Find out the effectiveness of the drug against the disease. (The table value of χ^2 for 1 degree of freedom at 5% level of significance is 3.84)

	Infection	No Infection	Total
Administering the drug	**144**	**312**	**456**
Without administering the drug	**192**	**72**	**264**
Total	**336**	**384**	**720**

Solution:

Null Hypothesis H_0: The drug is independent.

Alternative Hypothesis H_1: The drug is not independent

The expected frequencies are

$\frac{336 \times 456}{720} = 212.8$ ≈213	$\frac{384 \times 456}{720} = 243.2$ ≈243	456
$\frac{336 \times 264}{720} = 123.2$ ≈123	$\frac{384 \times 264}{720} = 140.8$ ≈141	264
336	384	720

O	E	O – E	$(O-E)^2$	$\left[\frac{(O-E)^2}{E}\right]$
144	213	- 69	4761	22.35
192	123	69	4761	38.71
312	243	69	4761	19.59
72	141	- 69	4761	33.77
				114.42

Calculated $\chi^2 = \left[\frac{(O-E)^2}{E}\right] = 114.42$

Degree of freedom = (r – 1) (c – 1)= (2-1)(2-1) = 1

Tabulated value of χ^2 at 5% level of significance with 1 degree of freedom is 3.841

Tabulated value = 3.841

Calculated value > Tabulated value, H_0 is rejected.

Therefore, the drug is definitely effective in controlling the typhoid.

21. A brand Manager is concerned that her brand's share may be unevenly distributed throughout the country. In a survey in which the country was divided into four geographical regions, a random sampling of 100 consumers in each region was surveyed, with the following results:

	Region				
	NE	**NW**	**SE**	**SW**	**TOTAL**
Purchased the brand	**40**	**55**	**45**	**50**	**190**
Did not purchase	**60**	**45**	**55**	**50**	**210**

Using χ^2 test, find out if the brand is unevenly distributed throughout the country.

Solution:

H_0: There is no significant difference between the observed and expected frequencies

H_1: There is a significant difference between the observed and expected frequencies

The expected frequencies are :

	Region				
	NE	NW	SE	SW	TOTAL
Purchased the brand	$\frac{190\times100}{400}\approx47$	$\frac{190\times100}{400}\approx48$	$\frac{190\times100}{400}\approx47$	$\frac{190\times100}{400}\approx48$	190
Did not purchase	$\frac{210\times100}{400}\approx53$	$\frac{210\times100}{400}\approx52$	$\frac{210\times100}{400}\approx53$	$\frac{210\times100}{400}\approx52$	210
	100	100	100	100	400

O	E	O – E	$(O-E)^2$	$\left[\frac{(O-E)^2}{E}\right]$
40	47	- 7	49	1.04
55	48	7	49	1.02
45	47	- 2	4	0.085
50	48	2	4	0.083
60	53	7	49	0.924
45	52	-7	49	0.942
55	53	2	4	0.075
50	52	- 2	4	0.076
				4.245

Calculated $\chi^2=\left[\frac{(O-E)^2}{E}\right]=4.245$

Degree of freedom = (r – 1)(c – 1)= (2-1)(4-1) = 3

Tabulated value of χ^2 at 5% level of significance with 3 degree of freedom is 7.815

Tabulated value = 7.815

Calculated value < Tabulated value, H_0 is accepted

There is no significant difference between the observed and expected frequencies.

TEXT / REFERENCE BOOKS

1. S. C. Gupta & V. K. Kapoor, Fundamental of Mathematical Statistics, 9th Edition, Sultan Chand & Sons, New Delhi, 1994.
2. P. R. Vittal, Mathematical Statistics, Margham Publications, Chennai, 2002.
3. Treatment and Content as in Mathematical Statistics, J. N. Kapur and H. C. Saxena, 20th Edition, S. Chand & Co. Ltd., New Delhi, 2010.
4. Hogg, R.V. & Craig. A. T. (1998): Introduction to Mathematical Statistics, Macmillan
5. T.Veerarajan, Probability, Statistics and Random process, Tata McGraw Hill, 1st reprint, 2004.