Computer Oriented Statistical Methods (A8005)

#1.1 PROBABILITY

In this section we shall explain the various terms which are used in the definition of probability under different approaches.

1.Random Experiment: If in each trial of an experiment conducted under identical conditions, the outcome is not unique, but may be any one of the possible outcomes, then such an experiment is called a random experiment.

Examples: Tossing a coin, throwing a die, selecting a card from a pack of playing cards, selecting a family out of a given group of families.

2. Outcome: The result of a random experiment will be called an outcome

3. Trial and Event: Any particular performance of a random experiment is called a trial and outcome or combination of outcomes is termed as events.

Example: If a coin is tossed repeatedly, the result is not unique. We may get any of the two faces, head or tail. Thus tossing a coin is random experiment or trial and getting of a head or tail is an event.

4. Exhaustive Events or Cases: The total number of possible outcomes of random experiment is known as the exhaustive events.

Example: In tossing coin there are two exhaustive cases (Head, Tail)

5.Favourable Events: The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of the event.

Example: In throwing of two dice, the number of cases favourable to getting the sum 5 is
(1,4), (4,1), (2,3), (3,2)

6. Mutually Exclusive Events: Events are said to be mutually exclusive or incompatible if the happening of any one of them precludes the happening of all the others, i.e., if no two or more of them can happen simultaneously in the same trial.

Example: In tossing a coin the events head and tail are mutually exclusive.

7. Equally Likely Events: Outcomes of trial are said to be equally likely if taking into consideration all the relevant evidences, there is no reason to expect one in preference to the others.

Example: In throwing an unbiased die, all the six faces are equally likely to come.

8. Independent Events: Several events are said to be independent if the happening of an event is not affected by the supplementary knowledge concerning the occurrence of any number of the remaining events.

Example: In tossing an unbiased coin, the event of getting a head in the first toss is independent of getting a head in the second, third and subsequent throws.

Sample Space: In probability theory, the sample space of an experiment or random trial is the set of all possible outcomes or results of that experiment.

Counting sample points: There are many counting techniques which can be used to count the number points in the sample space (or in some events) without listing each element. In many cases, we can compute the probability of an event by using the counting techniques.

Definition of Probability: If a random experiment or a trial results in 'n' exhaustive mutually exclusive and equally likely outcomes (or cases), out of which m are favourable to the occurrence of an event E, then the probability 'P' of occurrence (or happening) of E, usually denoted by $P(E)$ is given by :

$$P = P(E) = \frac{Favourable\ number\ of\ cases}{Exhaustive\ number\ of\ cases} = \frac{m}{n}$$

Axioms of Probability:

- **Axiom 1:** The probability of an event is a real number greater than or equal to 0.
- **Axiom 2:** The probability that at least one of all the possible outcomes of a process (such as rolling a die) will occur is 1.
- **Axiom 3:** If two events *A* and *B* are mutually exclusive, then the probability of either *A* or *B* occurring is the probability of *A* occurring plus the probability of *B* occurring.

Examples 1: Find the probability of getting a numbered card when a card is drawn from the pack of 52 cards.

Solution: Total Cards = 52.

Numbered Cards = (2, 3, 4, 5, 6, 7, 8, 9, 10)

9 from each suit 4 × 9 = 36

$$P(E) = \frac{36}{52} = \frac{9}{13}$$

Example 2: What is the probability of getting a sum of 7 when two dice are thrown?

Solution: Total number of ways = 6 × 6 = 36 ways.

Favorable cases = (1, 6) (6, 1) (2, 5) (5, 2) (3, 4) (4, 3) --- 6 ways.

$$P\ (A)\ =\ \frac{6}{36} = \frac{1}{6}$$

Example3: Two cards are drawn from the pack of 52 cards. Find the probability that both are diamonds or both are kings.

Solution: Total no. of ways = $52C_2$

Case I: Both are diamonds = $13C_2$

Case II: Both are kings = $4C_2$

P (both are diamonds or both are kings)= $\dfrac{13C_2 + 4C_2}{52C_2}$

Example 4: What is the probability of getting a sum of 22 or more when four dice are thrown?

Solution: Total number of ways = 6^4 = 1296.

Number of ways of getting a sum 22 are 6,6,6,4 = 4! / 3! = 4 and 6,6,5,5 = 4! / 2!2! = 6.

Number of ways of getting a sum 23 is 6,6,6,5 = 4! / 3! = 4.

Number of ways of getting a sum 24 is 6,6,6,6 = 1.

Favourable Number of cases = 4 + 6 + 4 + 1 = 15 ways.

P (getting a sum of 22 or more) = $\dfrac{15}{1296} = \dfrac{5}{432}$

Example 5: From a pack of cards, three cards are drawn at random. Find the probability that each card is from different suit.

Solution: Total number of cases =$52C_3$

One card each should be selected from a different suit.

The three suits can be chosen in $4C_3$ ways

The cards can be selected in a total of $4C_3$ x$13c_1$ x $13c_1$ x $13c_1$ ways

The required probability = $4C_3$ x $(13C_1)^3/52C_3$= 4 x $(13)^3$ / $52C_3$

Example 6: Find the probability that a leap year has 52 Sundays.

Solution: 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.

i.e., 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)

Exercise 1:

1. In a single throw with two dice find the probability of throwing a sum (i) 10 (ii) Which is perfect square [Ans: 1/12, 7/36]

2. What is the probability for a leap year to have 52 Mondays and 53 Sundays? [Ans: 1/7]

3. Five-digit numbers are formed with 0,1,2,3,4 (not allowing a digit being repeated in any number). Find the probability of getting 2 in the ten's place and 0 in the units place always. [Ans: 1/16]

4. In a class there are 10 boys and 5 girls. A committee of 4 students is to be selected from the class.Find the probability for the committee to contain at least 3 girls. [Ans: 0.0769]

5. A class consists of 6 girls and 10 boys. If a committee of 3 is chosen at random from the class, find the probability that (i) 3 boys are selected (ii) exactly 2 girls are selected. [Ans: 0.2143, 0.2678]

6. What is the probability that 4S's appear consecutively in the word MISSISSIPPI assuming that the letters are arranged at random. [Ans: 0.02424]

7. A bag contains 50 tickets numbered 1,2, 3, …., 50 of which five are drawn at random and arranged in ascending order of the magnitude. What is the probability that the middle one is 30? [Ans: 0.0364]

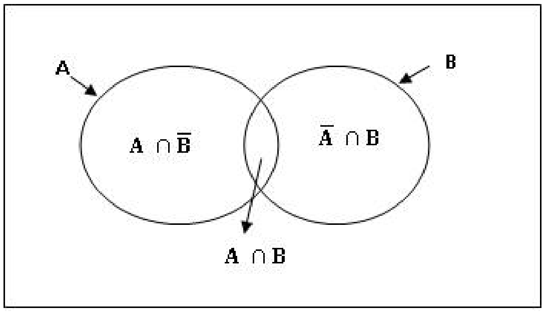
8. What is the probability that at least 2 out of 10 persons have the same birthday? Assume 365 days and that all days are equally likely.

Addition Theorem of probability:

Theorem 1: If A and B are any two events (subsets of sample space S) and are not disjoint, then

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

Proof:



From above venn diagram, we have $A\cup B=A\cup(\bar{A}\cap B)$

Where A and $\bar{A}\cap B$ are mutually disjoint.

$$\therefore\ P(A\cup B)=P(A\cup(\bar{A}\cap B)\ = P(A)+(\bar{A}\cap B)= P(A)+P(B)-(A\cap B)$$

Theorem 2: For three non- mutually exclusive events A, B and C, we have

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

Proof:

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

Examples on Addition Theorem for Mutually Exclusive Events

Example 1: A card is drawn at random from a pack of 52 cards. Find the probability that the drawn card is either a club or an ace of diamond.

Solution: Let A: Event of drawing a card of club and
B: Event of drawing an ace of diamond

The probability of drawing a card of club is, $P(A)=\frac{13}{52}$

The probability of drawing an ace of diamond is, $P(B)=\frac{1}{52}$

Since the events are mutually exclusive,

P(drawn card being a club or an ace of diamond)=

$$P(A\cup B)=P(A)+P(B)=\frac{13}{52}+\frac{1}{52}=\frac{14}{52}=\frac{7}{26}$$

Example 2: A box contains 30 chocolates numbered from 1 to 30. One chocolate is selected at random. Find the probability that number of the selected chocolate is a multiple of 5 or 8.

Solution: Let A be the event of number being a multiple of 5 within 30 and B be the event of number being a multiple of 8 within 30.

Favorable cases for event A are {5, 10, 15, 20, 25, 30}

Similarly, favorable cases for event B are {8, 16, 24}

The probability of the number being a multiple of 5 within 30 is $P(A)=\frac{6}{30}$

The probability of the number being a multiple of 8 within 30 is $P(B)=\frac{3}{30}$

Since A and B are mutually exclusive, the probability that number of the chocolate is a multiple of 5 or 8 is:

$$P(A\cup B)=P(A)+P(B)=\frac{6}{30}+\frac{3}{30}=\frac{9}{30}=\frac{3}{30}$$

Examples on Addition Theorem for Non-mutually Exclusive Events

Example 3. A card is drawn at random from a pack of 52 cards. Find the probability that the drawn card is either a spade or a king.

Solution: Let A be the event of drawing a card of spade and
B be the event of drawing a king card

The probability of drawing a card of spade is, $P(A)=\frac{13}{52}$

The probability of drawing a king card is, $P(B)=\frac{4}{52}$

Because one of the kings is a spade card also therefore these events are not mutually exclusive.

The probability of drawing a king of spade is, $P(A\cap B)=\frac{1}{52}$

The probability of the drawing a spade or king card,

$$= P(A\cup B)=P(A)+P(B)-P(A\cap B)=\frac{13}{52}+\frac{4}{52}-\frac{1}{52}=\frac{16}{52}=\frac{4}{13}$$

Example 4. A box contains 30 chocolates numbered from 1 to 30. One chocolate is selected at random. Find the probability that number of the selected chocolate is a multiple of 5 or 6.

Solution: Let A be the event of number being a multiple of 5 within 30 and
B be the event of number being a multiple of 6 within 30.
Favourable cases for event A are {5, 10, 15, 20, 25, 30}
Similarly, favourable cases for event B are {6, 12, 18, 24, 30}

The probability of the number being a multiple of 5 within 30 is $P(A)=\frac{6}{30}$

The probability of the number being a multiple of 6 within 30 is $P(B)=\frac{5}{30}$

Since 30 is a multiple of 5 as well as 6, therefore the events are not mutually exclusive is

$$P(A\cap B)=\frac{1}{30}$$

The probability that the number of the selected chocolate is a multiple of 5 or 6 is:

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)=\frac{6}{30}+\frac{5}{30}-\frac{1}{30}=\frac{10}{30}=\frac{1}{10}$$

Example 5. A number was drawn at random from the number 1 to 50. What is the probability that it will be a multiple of 2 or 3 or 10?

Solution: Probability of getting a multiple of 2: $P(A)=\frac{25}{50}$

Probability of getting a multiple of 3: $P(B)=\frac{16}{50}$

Probability of getting a multiple of 10: $P(C)=\frac{5}{50}$

Common Probability of getting a multiple of 2 and 3: $P(A\cap B)=\frac{8}{50}$

Common Probability of getting a multiple of 3 and 10: $P(B\cap C)=\frac{1}{50}$

Common Probability of getting a multiple of 2 and 10: $P(A\cap C)=\frac{5}{50}$

Common Probability of getting a multiple of 2, 3 and 10: $P(A\cap B\cap C)=\frac{1}{50}$

Probability that it is a multiple of 2 or 3 or 10:
$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(B\cap C)-P(A\cap C)+P(A\cap B\cap C)$

$$=\frac{25}{50}+\frac{16}{50}+\frac{5}{50}-\frac{8}{50}-\frac{1}{50}-\frac{1}{50}+\frac{1}{50}=\frac{33}{50}$$

Exercise 2:

1. Two dice are tossed. Find the probability of getting 'an even number on the first die or a total of 8. [Ans: 5/9]

2. An integer is chosen at random from two hundred digits. What is the probability that the integer is divisible by 6 or 8? [Ans: 1/4]

3. The probability that a student passes a Physics test is 2/3 and the probability that he passes both a Physics test and English test is 14/45. The probability that he passes at least one test is 4/5. What is the probability that he passes the English test? [Ans: 4/9]

4. Three newspapers A, B and C are published in a certain city. It is estimated from a survey that of the adult population: 20% read A, 16%read B, 14% read C, 8% read both A and B, 5% read both A and C, 4% read both B and C, 2% read all three. Find what percentage read at least one of the papers? [Ans: 2/100]

5. A card is drawn from a pack of 52 cards. Find and probability of getting a king or a heart or a red card. [Ans: 2/52]

6. 3 students A, B and C are in a running race. A and B have same probability of winning and each is twice as likely as C. Find (i) $P\left(B\ or\ C\right)$ wins (ii) P(A) looses.

[Ans: 3/5, 3/5]

Odds in favor and odds against an event:

Let P(A) and P(A^C) be happening and not happening of an event A.

If $P(A):P(A^c)$, then odds are in favor of the event

If $P(A^c):P(A)$, then odds are against the event

If $P(A):P(A^c)=a:b$ then $P(A)=\dfrac{a}{a+b}, P(A^c)=\dfrac{b}{a+b}$

Example 1: Fifteen people sit around a circular table. What are odds against two particular people sitting together?

Solution: 15 persons can be seated in 14! Ways.

No. of ways in which two particular people sit together is 13! × 2!

The probability of two particular persons sitting together $=\dfrac{13!2!}{14!}=\dfrac{1}{7}$

Odds against the event = 6 : 1

Example 2: An investment consultant predicts the odds against the price of a certain stock will go up during the next week is 2:1 and the odds in favor of the price remaining the same are 1:3. What is the probability of the price of the stock will go down during the next week?

Solution: Let A be the event that stock price will go up
and B be the event that stock price will remain same

Given $P(A^c):P(A)=2:1 \Rightarrow P(A)=\frac{1}{2+1}=\frac{1}{3}$

$P(B):P(B^c)=1:3 \Rightarrow P(B)=\frac{1}{1+3}=\frac{1}{4}$

P (that stock price grows up or remain same)

$= P(A\cup B)=P(A)+P(B)=\frac{1}{3}+\frac{1}{4}=\frac{7}{12}$

P (that stock price will go down)

$= P\left(A^c\cap B^c\right)=1-P(A\cup B)=1-\frac{7}{12}=\frac{5}{12}$

Independent Events: Several events are said to be independent if the happening of an event is not affected by the supplementary knowledge concerning the occurrence of any number of the remaining events.

Product rule for Independent events: The product rule is $P(E_1 \cap E_2) = P(E_1).P(E_2)$ where E_1 and E_2 are events that are independent. Explain that independence means that one event occurring has no effect on the probability of the other event occurring.

Example1: The probabilities that students A, B, C, D solve a problem are $\frac{1}{3}, \frac{2}{5}, \frac{1}{5}$ and $\frac{1}{4}$ respectively. If all of them try to solve the problem, what is the probability that the problem is solved.

Solution: Given the probability of A, B, C, D solving the problem is

$$P(A) = \frac{1}{3}, P(B) = \frac{2}{5}, P(C) = \frac{1}{5}, P(D) = \frac{1}{4}$$

The probability that the problem is not solved by A, B, C, D are

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

The probability that the problem is not solved when A, B, C, D try together (Independently)

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

$$= P(\overline{A}).P(\overline{B}).P(\overline{C}).P(\overline{D})$$

$$= \frac{2}{3}.\frac{3}{5}.\frac{4}{5}.\frac{3}{4} = \frac{6}{25}$$

$\therefore$ The probability that the problem is solved = $1 - \frac{6}{25} = \frac{19}{25}$

Example 2: A 6-sided fair die is rolled twice. What is the probability that both rolls have a result of 6?

Solution: It is important to establish that each die roll is independent. That is, if the first die roll result is 6, it will not affect the probability of the second die roll resulting in 6.

For an individual die roll, the P (rolling 6 on a die) = $\frac{1}{6}$

P (1st roll is 6 and 2nd roll is 6) = $\frac{1}{6} \text{x} \frac{1}{6} = \frac{1}{36}$ (By Product rule)

The probability that both die rolls are 6 is $\frac{1}{36}$

Exercise 3:

1. The odds that person X speaks the truth are 3:2 and the odds that person Y speaks the truth are 5:3 In what percentage of cases are they likely to contradict each other on an identical point. [Ans: 47.5%]

2. One shot is fired from each of the three guns. E_1, E_2, E_3 denote the events that the target is hit by the first, second and third guns respectively. If $P(E_1)=0.5, P(E_2)=0.5$ and $P(E_3)=0.8$ and E_1, E_2, E_3 are independent events, find the probability that
   (a)Exactly one hit is registered (b) at least two hits are registered. [Ans: 0.26, 0.7]

3. A and B throw alternatively with a pair of fair dice. A wins if he throws a sum of six before B throws a sum o seven. While B wins if he throws a sum of seven before A throws a sum of six. If A begins, show that his probability of winning is 30/61.

4. A can hit a target 3 times in 5 shots, B hits target 2 times in 5 shots and C can hit 3times in 4 shots. Find the probability of the target being hit when all them try. [Ans: 47/50]

5. The students in a class are selected at random one after the other for an examination. Find the probability that the boys and girls sit alternatively if there are
   (i) 4 boys and 3 girls (ii) 3 boys and 3 girls [Ans: 1/35, 1/10]

Conditional Probability: If E_1 and E_2 are two events in a sample space S and $P(E_1)\neq 0$, then the probability of E_2, after the event E_1 has occurred, is called the conditional probability of the event E_1 has occurred, is called the conditional probability of the event of E_2 given E_1 and is denoted by $P\left(\frac{E_2}{E_1}\right)$ or $P(E_2/E_1)$

and we define $P\left(\frac{E_2}{E_1}\right)=\frac{P(E_1\cap E_2)}{P(E_1)}$

Similarly, $P\left(\frac{E_1}{E_2}\right)=\frac{P(E_1\cap E_2)}{P(E_2)}$

Multiplication theorem of Probability:

For two events A and B,

$$P(A\cap B)=P(A).P(B/A),\quad P(A)>0$$
$$=P(B).P(A/B),\quad P(B)>0$$

Where $P(B/A)$ represents conditional probability of occurrence of B when the event A has already happened and $P(A/B)$ is the conditional probability of happening of A, given that B has already happened.

Important Formulas:

1. $P(\overline{A}\cup\overline{B})=P(\overline{A\cap B})=1-P(A\cap B)$
2. $P(\overline{A}\cap\overline{B})=P(\overline{A\cup B})=1-P(A\cup B)$
3. $P(A\cap\overline{B})=P(A)-P(A\cap B)$
4. $P(\overline{A}\cap B)=P(B)-P(A\cap B)$
5. $P(\overline{A}\cup B)=P(\overline{A})+P(B)-P(A\cap B)$
6. $P(\overline{B})=1-P(B)$

Example 1: If $P(A)=\frac{1}{3}, P(B)=\frac{1}{4}, P(A\cup B)=\frac{1}{2}$. Determine (i) $P\left(\frac{B}{A}\right)$ (ii) $P\left(\frac{A}{B^C}\right)$.

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

Now $P(A\cap B)=P(A)+P(B)-P(A\cup B)\quad=\frac{1}{3}+\frac{1}{4}-\frac{1}{2}=\frac{1}{12}$

(i) $P\left(\frac{B}{A}\right)=\frac{P(A\cap B)}{P(A)}=\frac{1/12}{1/3}=\frac{1}{4}$

(ii) $P(B^C)=1-P(B)=1-\frac{1}{4}=\frac{3}{4}$

$$P(A\cap B^C)=P(A)-P(A\cap B)=\frac{1}{3}-\frac{1}{12}=\frac{1}{4}$$

$$\therefore P\left(\frac{A}{B^C}\right)=\frac{P(A\cap B^C)}{P(B^C)}=\frac{1/4}{3/4}=\frac{1}{3}$$

Example 2: A bag contains 3 pink candies and 7 green candies. Two candies are taken out from the bag with replacement. Find the probability that both candies are pink.

Solution: Let A = event that first candy is pink and B = event that second candy is pink.

P (first candy is pink) $= P(A)=\frac{3}{10}$

Since the candies are taken out with replacement, this implies that the given events A and B are independent.

$$P\ (B/A)\ = P\ (B) = \frac{3}{10}$$

Hence by the multiplication law we get,

$$P\ (A\cap B)\ =\ P\ (A)\ .\ P\ (B/A)\ =\ \frac{3}{10}\times\frac{3}{10}=\frac{9}{100} = 0.09$$

Example 3: A bag has 4 white cards and 5 blue cards. We draw two cards from the bag one by one without replacement. Find the probability of getting both cards white.

Solution: Let A = event that first card is white and B = event that second card is white.

P (first card is white) $= P\ (A)\ = \frac{4}{9}$

Now $P\ (B)\ =\ P\ (B/A)=\frac{3}{8}$ (because the events given are dependent on each other)

So, $P\ (A\ \cap B)\ =\ P\ (A)\ .\ P\ (B/A)=\frac{4}{9}\times\frac{3}{8}\ =\frac{1}{6}$

Example 4: Two marbles are drawn in succession from a box containing 10 red, 30 white, 20 blue and 15 orange marbles, with replacement being made after each draw. Find the probability that

(i) Both are white (ii) First is red and second is white

Solution: Total no. Of marbles in the box=75

(i) Let E_1 be the event of drawing the first marble as white. Then $P(E_1)=\frac{30}{75}$

Let E_2 be the event of drawing second marble also white. Then $P(E_2)=\frac{30}{75}$

The probability that both marbles are white (with replacement)

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

(ii) Let E$_1$ be the event of drawing the first marble as red. Then $P(E_1) = \frac{10}{75} = \frac{2}{15}$

Let E$_2$ be the event of drawing second marble as white, if first is red.

Then $P\left(\frac{E_2}{E_1}\right) = \frac{30}{75} = \frac{2}{5}$

∴ The probability that the first marble is red and second is white

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

Exercise 4:

1. A bag contains 10 gold and 8 silver coins. Two successive drawings of 4 coins are made such that (i) coins are replaced before the second trial (ii) coins are not replaced before the second trial. Determine the probability that the first drawing will give 4 gold and the second 4 silver coins. [Ans: 49/31212, 35/7293]

2. An urn contains 20 red and 10 blue balls. Two balls are drawn from a bag one after the other without replacement. What is the probability that both the balls drawn are red? [Ans: 38/87]

3. Sixty percent of the employees of XYZ Corporation are college graduates. Of these ten percent are in sales. Of the employees who did not graduate from college, eight percent in the sales. What is the probability that (i) an employee selected at random is in scale?
   (ii) an employee selected at random is neither in scale nor a scale college graduate?
   [Ans: 0.38, 0.08]

4. In a certain town 40% have brown hair, 25% have brown eyes and 15% have both brown hair and brown eyes. A person is selected at random from the town.
   (i) if he has brown hair, what is the probability that he has brown eyes also?
   (ii) If he has brown eyes, determine the probability that he does not have brown hair?
   [Ans: 3/8, 2/5]

5. From a city population, the probability of selecting (i) a male or a smoker is 7/10, (ii)a male smoker is 2/5, and (iii) a male, if a smoker is already selected is 2/3. Find the probability of selecting (a) a non-smoker, (b) a male, and (c) a smoker, if a male is first selected.
   [Ans: 2/5, ½, 4/5]

Bayes' Theorem: Let E$_1$, E$_2$ ...., En be n mutually disjoint events with non-zero probabilities i.e., $P(E_i) \neq 0$,(i=1,2,3,.....n) of a random experiment. If A is any arbitrary event of the same experiment with $P(A) > 0$, then $P(E_i / A) = \dfrac{P(E_i)P(A/E_i)}{P(A)} = \dfrac{P(E_i)P(A/E_i)}{\sum_{i=1}^{n} P(E_i)P(A/E_i)}$, (i=1, 2, ..., n)

Proof: Let S be the sample space of the random experiment,

then $S = E_1 \cup E_2 \cup ...... \cup E_n$

We have $A = S \cap A = [E_1 \cup E_2 \cup ...... \cup E_n] \cap A$

where $E_1 \cap A, E_2 \cap A, ..., E_n \cap A$ are mutually disjoint events.

$$P(A) = P(E_1 \cap A) + P(E_2 \cap A) + ... + P(E_n \cap A)$$
$$= P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + ... + P(E_n)P(A/E_n)$$
$$= \sum_{i=1}^{n} P(E_i)P(A/E_i) \quad \text{------(1)}$$

By multiplication theorem of probability $P(A \cap E_i) = P(E_i)P(A/E_i)$ ------------(2)

Also $P(E_i / A) = \dfrac{P(A \cap E_i)}{P(A)} = \dfrac{P(E_i)P(A/E_i)}{\sum_{i=1}^{n} P(E_i)P(A/E_i)}$ [From (1) and (2)]

Note: The total probability, $P(A) = \sum_{i=1}^{n} P(E_i)P(A/E_i)$

Example1: Police plan to enforce speed limits by using radar traps at 4 different locations within the city limits. The radar traps at each of these locations L_1, L_2, L_3, L_4 are operated for 40%, 30%, 20% and 30% of the time. If a person who is speeding on his way to work has probabilities of 0.2, 0.1, 0.5 and 0.2 respectively of passing through these locations, what is the probability that he will be fined for over speed.

Solution: L_1, L_2, L_3, L_4 are locations where radar traps are operated respectively.

$$P(L_1) = 0.4, P(L_2) = 0.3, P(L_3) = 0.2, P(L_4) = 0.3$$

A: event of passing through locations and caught by radar traps

$$P(A/L_1) = 0.2, P(A/L_2) = 0.1, P(A/L_3) = 0.5, P(A/L_4) = 0.2$$

By total probability, P(that he will be fined for over speed)

$$P(A) = \sum_{i=1}^{n} P(L_i)P(A/L_i)$$
$$= P(L_1)P(A/L_1) + P(L_2)P(A/L_2) + P(L_3)P(A/L_3) + P(L_4)P(A/L_4)$$
$$= 0.27$$

Example2: Box I contain 10 white and 3 black marbles while Box II contains 3 white and 5 black marbles. Two marbles are drawn at random from Box I and placed in Box II. Then one marble is drawn at random from Box II. What is the probability that it is a white marble?

Solution: Let B_1 : event of drawing 2 white marbles from Box I

B_2 : event of drawing 2 black marbles from Box I and

B_3 : event of drawing 1 white and 1 black marbles from Box I

$$P\left(B_1\right)=\frac{{}^{10}C_2}{{}^{13}C_2},\ P\left(B_2\right)=\frac{{}^{3}C_2}{{}^{13}C_2},\ P\left(B_3\right)=\frac{{}^{10}C_1\,{}^{3}C_1}{{}^{13}C_2}$$

A: drawing white marble from Box II (after transfer)

$$P\left(A/B_1\right)=\frac{5}{10},\ P\left(A/B_2\right)=\frac{3}{10},\ P\left(A/B_3\right)=\frac{4}{10}$$

By total probability, P(that it is a white marble)

$$P(A)=\sum_{i=1}^{n}P(B_i)P(A/B_i)$$

$$=P(B_1)P(A/B_1)+P(B_2)P(A/B_2)+P(B_3)P(A/B_3)$$

$$=\frac{15}{26}\cdot\frac{5}{10}+\frac{1}{26}\cdot\frac{3}{10}+\frac{5}{3}\cdot\frac{4}{10}=\frac{59}{130}$$

Example 3: The chance that doctor A will diagnose a disease x correctly is 60%. The chance that a patient will die by his treatment after correct diagnosis is 40% and the chance of death by wrong diagnosis is 70%. A patient of doctor A, who had disease x, died. What is the chance that his disease was diagnosed correctly.

Solution: Let E$_1$ be the event that "disease x is diagnosed correctly by doctor A" and

E$_2$ be the event that "A patient of doctor A who has disease x died".

Then $P(E_1)=\dfrac{60}{100}=0.6$

$$P\left(\frac{E_2}{E_1}\right)=\frac{40}{100}=0.4$$

Now $P(\bar{E}_1)=1-0.6=0.40$ and $P\left(\dfrac{E_2}{\bar{E}_1}\right)=\dfrac{70}{100}=0.7$

∴ By Baye's theorem

A patient of doctor A with disease x, died, P(that his disease was diagnosed correctly)

$$P\left(\frac{E_1}{E_2}\right)=\frac{P(E_1).P(E_2/E_1)}{P(E_1).P(E2/E_1)+P(\bar{E}_1).P(E2/\bar{E}_1)}$$

$$=\frac{0.6\times0.4}{0.6\times0.4+0.4\times0.7}=\frac{6}{13}$$

Exercise 5:

1. Of the three men, the chances that a politician, a business man or an academician will be appointed as a vice- chancellor (V.C) of a University are 0.5,0.3,0.2 respectively. Probability that research is promoted by these persons if they are appointed as V.C are 0.3, 0.7,0.8 respectively.

   (i) Determine the probability that research is promoted

   (ii) If research is promoted, what is the probability that V.C is an academician?

2. In a bolt factory machines A, B, C manufacture 20%, 30% and 50% of the total of their output and 6%, 3% and 2% are defective. A bolt is drawn at random and found to be defective. Find the probabilities that it is manufactured from (i) Machine A (ii) Machine B (iii) Machine C. [Ans: 12/31, 9/31, 10/31]

3. First box contains 2 black, 3 red, 1 white balls; second box contains 1 black, 1 red, 2 white balls and third box contains 5 black, 3 red, 4 white balls. Of these a box is selected at random. From it a red ball is randomly drawn. If the ball is red, find the probability that it is from second box. [Ans:1/4]

4. Three machines I, II, III produce identical items. Of their respective output 5%, 4% and 3% are faulty. On a certain day machine I has produced 25%, of total output, Machine II 30% and Machine III the remainder. If an item is selected at random and found to be faulty. What are the chances that it was produced by the machine with highest output? [Ans: 0.355]

5. The probability that X, Y, Z will be elected as president of a club are 0.3, 0.5 and 0.2 respectively. The probability that membership fees of the club are increased is 0.8 if X is elected president, is 0.1 if Y is elected and is 0.4 if Z is elected. What is the probability that there is increase in membership fee? [Ans: 0.37]

6. Suppose 5 men out of 100 and 25 women out of 10,000 are color blind. A color blind person is chosen at random. What is the probability of the person being a male. (Assume male and female to be in equal numbers) [Ans: 0.95]

7. In a group consisting of equal number of men and women. 10% of women and 45% of men are unemployed. If a person is selected at random from the group then find the probability that the person is an employee. [Ans: 29/40]

8. Bag I contains 4 white, 3 black marbles and bag II contains 3 white, 5 black marbles. One marble is drawn from the bag I and placed in the bag II. Determine the probability that a marble now drawn from bag II is black. [Ans: 38/63]

Computer Oriented Statistical Methods (A8005)
#1.2 Random Variables and Probability Distributions

Random Variables

Definition: A real variable X whose value is determined by the outcome of a random experiment is called a random variable. A random variable X can also be regarded as real – value function defined on the sample space x, f(x) is the probability of occurrence of the event represented by x.

Example: The sample space corresponding to tossing of two coins, S=(HH,HT,TH,TT)
After the performance of experiment, we count the number of tails and denote it by X. The first outcome HH has no Tail, so X=0. Similarly X=1, denote the outcomes HT or TH and X=2, represents the outcome TT.
Thus X takes the values 0,1,2 i.e., X=0,1,2

Types of Random Variables:

Random variables are two types:
(i) Discrete Random Variable (ii) Continuous Random Variable

Discrete Random Variable: A random variable X which can take only a finite number of discrete values in an interval of domain is called a discrete random variable.

Example: Tossing a coin, throwing a dice, the number of defectives in a sample of electric bulbs, the number of printing mistakes in each page of a book, the number of telephone calls received by the telephone operator, etc.,

Continuous Random variable: A random variable X which can take values continuously i.e., which takes all possible values in a given interval is called a continuous random variable.

Example: The height, age, time, temperature and weight of individuals

Probability Distribution Function:

Let X be random variable. Then the probability distribution function associated with X is defined as the probability that the outcome of an experiment will be one of the outcomes for which $X(s) \le x, x \in R$. That is, the function F(x) or $F_X(x)$ defined by
$F_X(x) = P(X \le x) = P\{s : X(s) \le x\}, -\infty < x < \infty$ is called the distribution function of X.

Properties of Distribution Function:

If F is the distribution function of a random variable X and if a<b, then

i. $P(a < X \le b) = F(b) - F(a)$

ii. $P(a \le X \le b) = P(X = b) + F(b) - F(a)$

iii. $P(a < X < b) = [F(b) - F(a)] - P(X = b)$

iv. $P(a \le X < b) = [F(b) - F(a)] - P(X = b) + P(X = a)$

Discrete Probability Distribution:

Probability distribution of a random variable is the set of its possible values together with their respective probabilities. Suppose X is a discrete random variable with possible outcomes (values) $x_1, x_2, x_3, \ldots$. The probability of each possible outcome is $P_i = P(X = x_i) = p(x_i)$ for i=1,2,3,....

If numbers $P(x_i)$, i=1, 2, 3, ... satisfy the two conditions

(i) $P(xi) > 0$ for all values of i; $0 < P_i < 1$

(ii) $\sum P(x_i) = 1$, i=1, 2, 3, ...

Then the function is $P(x)$ is called the probability mass function of random variable X and set $\{P(x_i)\}$, i=1, 2, ... is called the discrete probability distribution of the discrete random variable X.

Probability Density Function:

The probability density function is defined as the derivative of the probability distribution function. $F_X(x)$ of the random variable X.

Thus $f_X(x) = \dfrac{d}{dx}[F_X(x)]$

Mean and variance of Probability distribution:

The mathematical Expectation or mean or expected value of X, denoted by E(X), is defined as the sum of products values of x and corresponding probabilities.

$$\therefore E(X) = x_1 p_1 + x_2 p_2 + x_3 p_3 + \ldots\ldots x_n p_n$$

$$i.e E(X) = \sum_{i=1}^{n} p_i x_i \text{ Similarly } E(X^r) = \sum_{i=1}^{n} p_i x_i^r$$

Variance characterizes the variability in the distribution since two distributions with sane mean can still have different dispersion of data about their means.

Variance of the probability distribution of random variable X is

$$V(X) = E[X - E(X)]^2 \quad i.e., \quad Var[X] = \sum_{i=1}^{n} \{[xi - E(X)]^2 p_i(x_i)\}$$

Some Important result of variance:

1. Variance of constant is zero i.e., $V(K) = 0$
2. If K is constant, then $V(KX) = K^2 V(X)$
3. If X is discrete random variable, then $V(aX + b) = a^2 V(X)$, where $V(X)$ is variance of X and a,b are constants.

Example 1: A random variable X has the following probability function :

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

(i)Determine K (ii) Evaluate $P\left(X<6\right), P\left(X\geq 6\right), P\left(0<x<5\right)$ and $P\left(0\leq X\leq 4\right)$

(iii) If $P\left(X\leq K\right)>\frac{1}{2}$, find the minimum value of K and,

(iv) Determine the distribution function of X (v)Mean (vi)Variance.

Solution:

(i) Since $\sum_{x=0}^{7} p(x)=1$, we have

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

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

$$i.e(10K-1)(K+1)=0$$

$$\therefore K=\frac{1}{10}=0.1 \qquad \left(\text{since P}\left(\text{x}\right)\geq 0,\text{so K}\neq -1\right)$$

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

$$=0+K+2K+2K+3K+K^2=8K+K^2=0.8+0.01=0.81$$

$$P(X\geq 6)=1-P(X<6)=1-0.81=0.19$$

$$P\left(0<X<5\right)=P(X=1)+P(X=2)+P(X=3)+P(X=4)$$

$$=K+2K+2K+3K=8K=\frac{8}{10}=0.8$$

$$P(0\leq X\leq 4)=P(X=0)+P(X=1)+......+P(X=4)$$

$$=0+K+2K+2K+3K=8K=\frac{8}{10}=0.8$$

(iii) The required minimum value of K is obtained as below.

$$P(X\leq 1)=P(X=0)+P(X=1)=0+K=\frac{1}{10}=0.1$$

$$P(X \le 2) = P(X=0) + P(X=1) + P(X=2) = \frac{1}{10} + \frac{2}{10} = \frac{3}{10} = 0.3$$

$$P(X \le 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$$
$$= 0.3 + 0.2 = 0.5$$

$$P(X \le 4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)$$

=0.5+0.3=0.8>0.5=1/2

$\therefore$ The minimum value of K for which $P(X \le K) > \frac{1}{2}$ is K=4

(iv) The distribution of X is given by the following table:

X	0	1	2	3	4	5	6	7
P(x)	0	K	2K	2K	3K	K^2	$2K^2$	$7K^2$+K
F(x)	0	K=1/10	3K=3/10	5K=5/10	8K=8/10	8K+K^2=81/100	8K+3K^2=83/100	9K+10K^2=1

(v) Mean $\mu = \sum_{x=0}^{7} p_i x_i$ = 0(0) +1(k)+2(k)+3(2k)+4(3k)+5(k^2)+6(2k^2)+7(7K^2+k)

= 66K^2+30K=0.66+3=3.66

(vi) Variance = $\mu = \sum_{x=0}^{7} p_i x_i^2 - \mu^2$

$$= K + 8K + 18K + 48K + 25K^2 + 72K^2 + 343K^2 + 49K - (3.66)^2$$

=4.4+12.4-13.3956=3.4044

Example2: A sample of 4 items is selected at random from a box containing 12 items of which 5 are defective. Find the expected number E of defective items.

Solution: Let X denote the number of defective items among 4 items drawn from 12 items

X can take the values 0,1,2,3 or 4.

No. of good items=7 and No. of defective items=5

P(X=0)= P(no defective)= $\frac{7C_4}{12C_4} = \frac{7}{99}$, $P(X=1) = \frac{7C_3 \times 5C_1}{12C_4} = \frac{35}{99}$

$P(X=2) = \frac{7C_2 \times 5C_2}{12C_4} = \frac{42}{99}$, $P(X=3) = \frac{7C_1 \times 5C_3}{12C_4} = \frac{14}{99}$

$$P(X=4)=\frac{5C_4}{12C_4}=\frac{1}{99}$$

Discrete probability is

$X=x_i$	0	1	2	3	4
$P(X=x_i)$	7/99	35/99	42/99	14/99	1/99

Expected number of defective items= $E(X)=\sum xiP(X=x_i)=\frac{165}{99}$

Exercise 1:

1. Two dice are thrown. Let X assign to each point (a,b) in S the Maximum of its number i.e., X(a,b)=max(a,b). Find the probability distribution. X is a random variable with X(s)={0,1,2,3,4,5,6}. Also find the mean and variance of the distribution. [Ans:4.47, 1.989]

2. From a lot of 10 items containing 3 defectives, a sample if 4 items is are drawn at random. Let the random variable x denote the number of defective items in the sample. Find the probability distribution of X when the sample is drawn without replacement.

3. A random variable X has the following probability function

x	1	2	3	4	5	6	7	8
P(x)	K	2K	3K	4K	5K	6K	7K	8K

Find the value of (i)K (ii) P(X≤2)(iii) P2≤x≤5) [Ans: 1/36, 0.083, 0.3889]

4. For the following probability function, Find $E\left[X^2\right]$ and $E\left[(2X+1)^2\right]$

x	-3	6	9
P(x)	1/6	1/2	1/3

5. Given the following table:

X	-3	-2	-1	0	1	2	3
$P(x)$	0.05	0.10	0.30	0	0.30	0.15	0.10

Compute $(i)E(X)$ $(ii)E(2X\pm3)$ $(iii)E(4X+5)$ $(iv)E(X^2)$
$(v)V(X)$ $(vi)V(2X\pm3)$

6. A fair coin is tossed until a head or five tails occurs. Find Expected number E of tosses of the coin. [Ans: 1.9375]

7. A player tosses two fair coins. He wins Rs.100/- If head appears, Rs.200/- If two heads appear. On the other hand he losses Rs.500/- If no head appears, Determine the expected value E of the game and is the game favorable to the player? [Ans: -25, No]

Continuous Probability Distribution:

When a random variable X takes every value in an interval, it gives rise to continuous distribution of X. The distribution defined by variates like temperature, height and weights are continuous distribution.

Probability Density Function:

The probability density function is defined as the derivative of the probability distribution function. $F_X(x)$ of the random variable X.

Thus $f_X(x) = \frac{d}{dx}[F_X(x)]$

Properties of probability density function f(x)

(i) $f(x) \geq 0, \forall x \in R$

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

(iii) The probability P(E) is given by, $P(E) = \int_E f(x)dx$ is well defined for any event E.

Mean, Median, Mode, Variance of Probability distribution:

(i) Mean of a distribution is given by $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$

(ii) Median is the point which divides the entire distribution into two equal parts. In case of continuous distribution, median is the point which divides the total area into two equal parts. Thus if X is defined from a to b and M is the median, then

$$\int_a^M f(x)dx = \int_M^b f(x)dx = \frac{1}{2}$$

Solving for M, we get the median.

(iii) Mode is the value of X for which f(x) is maximum. Mode is thus given by $f'(x) = 0$ and $f''(x) < 0 \ for\ a < x < b$.

(iv) Variance of a distribution is given by $\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$

Cumulative distribution function of continuous random variable:

The continuous distribution function or simply the distribution function of a continuous random variable X is denoted by F(x) and is defined as

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

Thus F(x) gives the probability that the value of the variable X will be ≤x.

Example:1

The diameter, say X, of an electric cable, is assumed to be a continuous random variable with $p.d.f$: $f(x) = 6x(1-x)\ , 0 \le x \le 1$

(i) Check that the above is a p.d.f.,

(ii) Obtain an expression for the c.d.f. of X.,

(iii) Compute $P\left(X \le \frac{1}{2} / \frac{1}{3} \le X \le \frac{2}{3}\right)$, and

(iv) Determine the number k such that $P(X < K) = P(X > K)$

Solution: Since $\int_0^1 f(x)dx = \int_0^1 6x(1-x)dx = 6\left[\frac{x^2}{2} - \frac{x^3}{3}\right]_0^1 = 1$, $f(x)$ is a $p.d.f.$,

(ii) $$F(x) = \begin{cases} 0, if \ x \le 0 \\ (3x^2 - 2x^3), 0 < x \le 1 \\ 1, ifx > 1 \end{cases}$$

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

$$= \frac{\int_{1/3}^{1/2} 6x(1-x)dx}{\int_{1/3}^{2/3} 6x(1-x)dx} = \frac{11/54}{13/27} = \frac{11}{26}$$

(iii) We have $P(X < k) = P(X > k)$

$$\Rightarrow \int_0^k 6x(1-x)dx = \int_k^1 6x(1-x)dx$$

$$3k^2 - 2k^3 = 3(1-k^2) - 2(1-k^3) \Rightarrow 4k^3 - 6k^2 + 1 = 0 \Rightarrow k = \frac{1}{2}, \frac{1 \pm \sqrt{3}}{2}$$

The only admissible value of k in the given range is $\frac{1}{2}$. Hence the value of k is $\frac{1}{2}$.

Example 2: The probability density function $f(x)$ of a continuous random variable is given by $f(x) = ce^{-|x|}, -\infty < x < \infty$. Show that $c = \frac{1}{2}$ and find that the mean and variance of the distribution. Also find the probability that the variance lies between 0 and 4.

Solution: Given $f(x) = ce^{-|x|}, -\infty < x < \infty$

We have $\int\limits_{-\infty}^{\infty} f(x)dx = 1$ (since the total probability is unity)

$$\Rightarrow \int\limits_{-\infty}^{\infty} ce^{-|x|}dx = 1 \quad \Rightarrow c\int\limits_{-\infty}^{\infty} e^{-|x|}dx = 1$$

$$\Rightarrow 2c\int\limits_{0}^{\infty} e^{-|x|}dx = 1 \,[\, e^{-|x|} \text{ is an even function}] \qquad \Rightarrow 2c\int\limits_{0}^{\infty} e^{-x}dx = 1 \,[\because \text{ in } 0 \le x \le \infty, |x| = x\,]$$

$$\Rightarrow 2c(-e^{-x})_0^{\infty} = 1 \quad \Rightarrow -2c(0-1) = 1 \quad \Rightarrow 2c = 1 \quad \Rightarrow c = \frac{1}{2} \qquad \text{Hence } f(x) = \frac{1}{2}e^{-|x|}$$

(i) Mean of the distribution

$$\mu = \int\limits_{-\infty}^{\infty} xf(x)dx = \frac{1}{2}\int\limits_{-\infty}^{\infty} xe^{|-x|}dx = 0 \quad \text{[since integrant is odd.]}$$

(ii) Variance of the distribution

$$\sigma^2 = \int\limits_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int\limits_{-\infty}^{\infty} (x-0)^2 \frac{1}{2}e^{-|x|}dx = \frac{1}{2}\int\limits_{-\infty}^{\infty} x^2 e^{-|x|}dx$$

$$= 2.\frac{1}{2}\int\limits_{0}^{\infty} x^2 e^{-|x|}dx, \text{ since integrand is even}$$

$$= \int\limits_{0}^{\infty} x^2 e^{-|x|}dx = \left(x^2\frac{e^{-x}}{-1} - 2x\frac{e^{-x}}{1} + 2\frac{e^{-x}}{-1} \right)_0^{\infty} = [0-(-2)] = 2$$

(iii) The probability between 0 and 4 is

$$P(0 \le X \le 4) = \frac{1}{2}\int\limits_{0}^{4} e^{-|x|}dx = \frac{1}{2}\int\limits_{0}^{4} e^{-|x|}dx \quad [\because in 0 < x < 4, |x| = x\,]$$

$$= -\frac{1}{2}(1-e^{-4}) = -\frac{1}{2}(e^{-4}-1) = \frac{1}{2}(1-e^{-4}) = 0.4908 \text{ (nearly)}$$

Exercise 2:

1. A continuous random variable X has the distribution function

$$F(x) = \begin{cases} 0, \quad if \ x \le 1 \\ k(x-1)^4, if \ 1 < x \le 3 \\ 1, if \ x > 3 \end{cases}$$

Determine (i) $f(x)$ (ii) k (iii) Mean

2. If the probability density of a random variable is given by

$$f(x) = \begin{cases} k(1-x^2), for 0 < x < 1 \\ 0, otherwise \end{cases}$$

Find the value of k and the probability that a random variable having this probability density will take on a value (i) between 0.1 and 0.2 (ii) greater than 0.5.

3. If X is the continuous random variable whose density function is

$$f(x) = \begin{cases} x \quad if \ 0 < x < 1 \\ 2-x \ \ if 1 \le x < 2 \\ 0, \quad otherwise \end{cases}. \quad Find \ E(25X^2 + 30X - 5)$$

4. For continuous probability function $f(x) = kx^2 e^{-x}, x \ge 0.$ Find k, Mean and Variance.

[Ans: ½, 3, 3]

5. A r.v X gives measurements between 0 and 1 with a probability function

$$f(x) = \begin{cases} 12x^3 - 21x^2 + 10x, for \ 0 < x < 1 \\ 0, otherwise \end{cases}.$$

(i)Find $P(X \le 1/2)$ and $P(X > 1/2)$ $\quad (ii)$Find $k \ni P(X \le k) = 1/2$

6. If X is a c.r.v and $Y = aX + b.$. Prove that $E[Y] = aE[X] + b$ and $V[Y] = a^2 V[X]$, where a and b are constants.

Computer Oriented Statistical Methods (A8005)

#2 Discrete and Continuous Distributions

Distributions:

The distributions which are based on expectations on the basis of past experiences are theoretical distributions. The following distributions are discussed in this section:

Discrete theoretical distributions:

1. Binomial Distribution
2. Poisson Distribution

Continuous theoretical distributions:

1. Uniform Distribution
2. Normal Distribution

Bernoulli Distribution:

A random experiment with only two possible outcomes, success or failure is called Bernoulli trial. The corresponding distribution is Bernoulli distribution.

The probability function of Bernoulli's distribution is $p(x) = p^x q^{1-x} = p^x (1-p)^{1-x}, x = 0,1.$

Binomial Distribution:

A random variable X has a binomial distribution if it assumes only non-negative values and its probability density function is given by

$$P(X=x) = p(x) = \begin{cases} n_{C_x} p^x q^{n-x}; x = 0,1,2.....n; q = 1-p \\ 0, \text{otherwise} \end{cases} = b(x;n,p)$$

Physical Conditions for Binomial Distribution:

The distribution can be used under the following conditions

1. The number of trials is finite and fixed.
2. In every trial , there are only two possible outcomes—success or failure
3. The outcomes are independent. The outcome of one trial does not affect the other trial.
4. The P(Success) is same for each trial.

Examples of binomial distribution:

1. The no. of defective bolts in a box containing 'n' bolts.
2. The no. of machines lying idle in a factory having 'n' machines.
3. The no. of oil wells yielding natural gas in a group of 'n' wells test drilled.

Constants of Binomial Distribution

Mean of the Binomial Distribution:

The Binomial probability distribution is given by

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

Mean of X, $\mu = E(X) = \sum_{x=0}^{n} x\, p(x) = \sum_{x=0}^{n} x\, n_{C_x} p^x q^{n-x}$

$$=\sum_{x=0}^{n} x\frac{n!}{(n-x)!x!}p^x q^{n-x}=\sum_{x=1}^{n}\frac{(n-1)!n}{(n-x)!(x-1)!}\;p\;p^{x-1}q^{n-x}$$

$$=np\sum_{x=1}^{n}\frac{(n-1)!}{(n-x)!(x-1)!}\;\;p^{x-1}q^{n-x}=np\sum_{x=1}^{n}n-1_{C_{x-1}}\;\;p^{x-1}q^{n-x}$$

$$=np\left(p+q\right)^{n-1}=np\left(1\right)=np\quad\left(\because p+q=1\right)$$

Hence the Mean of Binomial Distribution $=np$

Variance of Binomial Distribution:

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

$$=\sum_{x=0}^{n}x^2\;p\left(x\right)-\mu^2=\sum_{x=0}^{n}\left(x\left(x-1\right)+x\right)p\left(x\right)-\mu^2$$

$$=\sum_{x=0}^{n}x\left(x-1\right)n_{C_x}p^x q^{n-x}+\sum_{x=0}^{n}xp\left(x\right)-\mu^2$$

$$=\sum_{x=0}^{n}x\left(x-1\right)\frac{n!}{(n-x)!x!}p^x q^{n-x}+\mu-\mu^2$$

$$=n(n-1)p^2\sum_{x=2}^{n}\frac{(n-2)!}{(n-x)!(x-2)!}\;\;p^{x-2}q^{n-x}+\mu-\mu^2$$

$$=n\left(n-1\right)p^2\left(p+q\right)^{n-2}+\mu-\mu^2=n\left(n-1\right)p^2\left(1\right)^{n-2}+\mu-\mu^2$$

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

$\therefore$ Variance of the Binomial distribution $=npq$

Hence the Standard Deviation of Binomial distribution $=\sqrt{npq}$

Example 1: Ten coins are thrown simultaneously. Find the probability of getting at least Seven Heads

Solution: p=Probability of getting a head $=\frac{1}{2}$, q= Probability of not getting a head $=\frac{1}{2}$

The probability of getting x heads in a throw of 10 coins is

$$P\left(X=x\right)=p\left(x\right)=10_{C_x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{10-x}\quad,x=0,1,2,.....,10$$

Probability of getting at least seven heads is given by

$$=P\left(X\geq 7\right)=P\left(X=7\right)+P\left(X=8\right)+P\left(X=9\right)+P\left(X=10\right)$$

$$=10C_7\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^3+10C_8\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2+10C_9\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right)^1+10C_{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^0$$

$$=\frac{1}{2^{10}}\left[10C_7+10C_8+10C_9+10C_{10}\right]=\frac{1}{2^{10}}[120+45+10+1]=\frac{176}{1024}=0.1719$$

Example 2: The mean and variance of binomial distribution are 4 and $\frac{4}{3}$ respectively. Find $P(x\geq 1)$.

Solution:

Mean of binomial distribution=np=4................................(1)

Variance binomial distribution =npq=$\frac{4}{3}$ ....(2)

$$\frac{(2)}{(1)}=\frac{npq}{np}=q=\frac{1}{3}$$

$$\therefore p=1-q=1-\frac{1}{3}=\frac{2}{3}$$

$$n=\frac{4}{p}=4\times\frac{3}{2}=6$$

$$\therefore p(X\geq 1)=1-p(X<1)=1-p(X=0)=1-6C_0\,p^0q^{6-0}$$

$$=1-1.\left(\frac{2}{3}\right)^0.\left(\frac{1}{3}\right)^6=1-\left(\frac{1}{3}\right)^6=0.999$$

Example 3: The probability of a man hitting a target is $\frac{1}{4}$.

(i) If he fires 7 times, what is the probability of his hitting the target (a) at least twice (b) at most 3 times?

(ii) How many times must he fires so that the probability of his hitting the target at least once is more than $\frac{2}{3}$?

Solution:

n= number of trials = 7 , p =The probability of hitting a target =$\frac{1}{4}$

q =The probability of not hitting a target=$1-\frac{1}{4}=\frac{3}{4}$

(i) (a) $p(\text{at least twice})=p(X\geq 2)=1-p(\text{X<2})=1-\left[p(\text{X=0})+p(\text{X=1})\right]$

$$=1-\left[7C_0\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^7+7C_1\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^{7-1}\right]=\frac{4547}{8192}=0.555$$

(b) $p(\text{at most 3 times}) = p(X \le 3) = p(\text{X=0}) + p(\text{X=1}) + p(\text{X=2}) + p(\text{X=3})$

$$= \sum_{x=0}^{3} {}^7C_x \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{7-x} = \frac{3807}{4096} = 0.929$$

(ii)Given that $P(\text{at least once}) > \frac{2}{3}$

$$P(\text{X} \ge 1) > \frac{2}{3} = 1 - P(\text{X} < 1) > \frac{2}{3} = 1 - P(\text{X} = 0) > \frac{2}{3}$$

$$i.e.,\ 1 - {}^nC_0 \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^n > \frac{2}{3} \Rightarrow 1 - \left(\frac{3}{4}\right)^n > \frac{2}{3} \Rightarrow \frac{1}{3} > \left(\frac{3}{4}\right)^n$$

This is satisfied for $n = 4$

Exercise 1:

1. Find max n ∋ probability of getting no head in tossing a coin n times >0.1 [Ans: n =3]

2. In a binomial distribution consisting of 5 independent trials, probabilities of 1 and 2 successes are 0.4096 and 0.2048 respectively. Find the parameter p of the distribution [Ans: $p = 0.2$]

3. A multiple-choice test consists of 8 questions with 3 answers to each question (of which only one is correct). A student answers each question by rolling a balanced die and checking the first answer if he gets 1 or 2, the second answer if he gets 3 or 4 and the third answer if he gets 5 or 6. To get a distinction, the student must secure at least 75% correct answers. If there is no negative marking, what is the probability that the student secures a distinction? [Ans: 0.0197]
[Hint: 75% correct answers implies 6 of 8 questions should be correct)

4. A coffee connoisseur claims that he can distinguish between a cup of instant coffee and a cup of percolator coffee 75% of the time. It is agreed that his claim will be accepted if he correctly identifies at least 5 of the 6 cups. Find his chances of having the claim (i) accepted, (ii) rejected, when he does have the ability he claims. [Ans: (i) 0.534, (ii) 0.466]

Examples of Binomial Distributions:

If n independent trials constitute one experiment and this experiment is repeated N times, the frequency distribution of x success is

$$f(x) = Np(x) = Nn_{C_x} p^x q^{n-x};\ x = 0,1,2,\ldots,n$$

Example 1: Six dice are thrown 729 times .How many times do you except at least three dice to show a 5 or 6 ?

Solution: p =Probability of occurrence of 5 or 6 in one throw$= \frac{2}{6} = \frac{1}{3}$

$\therefore\ q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}$ and $n = 6$

The probability of getting at least three dice to show a 5 or 6

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

$$= 6C_3\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^3 + 6C_4\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^2 + 6C_5\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^1 + 6C_6\left(\frac{1}{3}\right)^6\left(\frac{2}{3}\right)^0 = \frac{233}{729} = 0.319$$

$\therefore$ The expected number of such cases in 729 times = $N.p(x) = 729 \times 0.319 = 232.55$ □ 233

Example 2: Out of 800 families with 5 children each, how many would you expect to have
(a) 3 Boys (b) 5 Girls (c) Either 2 or 3 boys (d) at least one boy?
Assume equal probabilities for boys and girls.

Solution: No. of children = n =5

p =Probability of each boy $= \frac{1}{2}$ q = Probability of each girl$= \frac{1}{2}$

The probability distribution is

$$p(x) = {}^nC_x p^x q^{n-x} = {}^5C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{5-x} = \frac{1}{2^5}{}^5C_x \text{ Per family}$$

(i) $P(3\text{ boys}) = P(X=3) = p(3) = \frac{1}{2^5}{}^5C_3 = \frac{10}{32} = \frac{5}{16} = 0.3125$, per family

Thus for 800 families the probability of number of families having 3 boys
$= 0.3125 \times 800 = 250$ Families

(ii) $P(5\text{ girls}) = P(no\ boys) = P(X=0) = p(0) = \frac{1}{2^5}{}^5C_0 = \frac{1}{32} = 0.03125$, per family

Thus for 800 families the probability of number of families having 5 girls
$= 0.03125 \times 800 = 25$ Families

(iii) $P(\text{either 2 or 3 boys}) = P(X=2) + P(X=3) = p(2) + p(3)$

$$= \frac{1}{2^5}{}^5C_2 + \frac{1}{2^5}{}^5C_3 = \frac{1}{2^5}(10+10) = \frac{20}{32} = \frac{5}{8}$$

$\therefore$ Expected number of families with 2or 3 boys=$\frac{5}{8} \times 800 = 500$ families

(iv) $P\left(\text{at least one boy}\right) = P(X \geq 1) = 1 - \left[P(X=0)\right]$

$$= 1 - \frac{1}{2^5}(1) = \frac{31}{32} = 0.96875$$

$\therefore$ Expected number of families with one boy= $0.96875 \times 800 = 775$

Example 3: In a precision bombing attack there is a 50% chance that any one bomb will strike the target. Two direct hits are required to destroy the target completely. How many bombs must be dropped to give a 99% chance or better of completely destroying the target?

Solution: p =Probability that the bomb strikes the target =50% $= \frac{1}{2}$

Let n be the number of bombs which should be dropped to ensure 99% chance or better of completely destroying the target. This implies 'probability that out of n bombs, at least two strike the target, is greater than 0.99'.

Let X be a $r.v$ representing the number of bombs striking the target.

Then $P\left(X = x\right) = p(x) = n_{c_x}\left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{n-x} = n_{c_x}\left(\frac{1}{2}\right)^n ; x = 0,1,2,..,n$

We should have

$$P(X \geq 2) \geq 0.99 \qquad \Rightarrow \left[1 - P(X \leq 1)\right] \geq 0.99$$

$$\Rightarrow 1 - P(X=0) - P(X=1) \geq 0.99$$

$$\Rightarrow P(X=0) + P(X=1) \leq 0.01 \qquad \Rightarrow \frac{1}{2^n}(1+n) \leq 0.01$$

$$\Rightarrow \frac{1}{0.01}(1+n) \leq 2^n \qquad \Rightarrow 100(1+n) \leq 2^n$$

By trial method, $n = 11$. Hence the minimum number of bombs needed to destroy the target completely is 11.

Exercise 2:

1. In 256 sets of 12 tosses of a coin, in how many cases one can expect 8 heads and 4 tails. [Ans:31]

2. In sampling a large number of parts manufactured by a machine, the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples, how many would be expected to contain at least 3 defective parts. [Ans: 323]

3. An irregular six-faced die is thrown and the expectation that in 10 throws it will give five even numbers is twice the expectation that it will give four even numbers. How many items in 10,000 sets of 10 throws each, would you expect it to give no even number? [Ans: 1]

Fitting of Binomial Distribution:

Recurrence Relation for the probabilities of Binomial Distribution:

We have $\frac{p(x+1)}{p(x)} = \frac{n_{c_{x+1}} p^{x+1} q^{n-x-1}}{n_{c_x} p^x q^{n-x}} = \frac{n-x}{x+1}.\frac{p}{q}$ (on simplification)

$\therefore p(x+1) = \left(\frac{n-x}{x+1}\right).\frac{p}{q} p(x)$ is recurrence formula for the probabilities of Binomial Distribution.

Example 1: Seven coins are tossed and the no. of heads are noted. The experiment is repeated 128 times and the following distribution is obtained.

No.of Heads	0	1	2	3	4	5	6	7	Total
Frequency	7	6	19	35	30	23	7	1	128

Fit binomial distribution assuming (a) the coin is unbiased (b) the nature of the coin is not known.

Solution:

(a) The coin is unbiased

$\therefore\ p = \frac{1}{2},\ q = \frac{1}{2}$, n =7, N= $\sum f_i$ =7+6+19+35+30+

Using recurrence relation $p(x+1) = \frac{(n-x)p}{(x+1)q}.p(x)$

$$= \frac{7-x}{x+1}.p(x)\left[\because n = 7, \frac{p}{q} = 1\right]$$

$\therefore\ p(0) = 7_{c_0} p^0 q^7 = 7c_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^7 = \left(\frac{1}{2}\right)^7$ $\left(\text{Using } p(x) = n_{C_x} p^x q^{n-x}\right)$

No.of Heads (x)	Observed Frequency	Probability , $p(x)$	Expected or Theoretical Frequency $f(x) = N.p(x) = 128.p(x)$
0	7	$p(0) = \frac{1}{2^7}$	$f(0) = 128.p(0) = 128 \times \frac{1}{2^7} = 1$
1	6	$p(1) = 7.p(0) = 7/2^7$	$f(1) = 128.p(1) = 128 \times 7/2^7 = 7$
2	10	$p(2) = 3.p(1) = 21/2^7$	$f(2) = 128.p(2) = 128 \times 21/2^7 = 21$
3	35	$p(3) = \frac{35}{2^7}$	$f(3) = 128.p(3) = 35$
4	30	$p(4) = \frac{35}{2^7}$	$f(4) = 128.p(4) = 35$
5	23	$p(5) = \frac{21}{2^7}$	$f(5) = 128.p(5) = 21$
6	7	$p(6) = \frac{7}{2^7}$	$f(6) = 128.p(6) = 7$
7	1	$p(7) = \frac{1}{2^7}$	$f(7) = 128.p(7) = 1$

(b) The nature of the coin is not known.

We are given n=7 and $N = \Sigma f_1 = 128$

$$\therefore \text{ Mean } = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{6+38+105+120+115+42+7}{128} = \frac{433}{128} = 3.383$$

i.e., $np = 3.383 \Rightarrow 7p = 3.383$

$$\therefore p = \frac{3.383}{7} = 0.4833 \text{ and } q = 1 - p = 1 - 0.4833 = 0.5167$$

By binomial distribution $p(x) = {}^{n}C_{x}\, p^{x} q^{n-x} = {}^{7}C_{x} (0.4833)^{x} (0.5167)^{7-x}$

No.of Heads (x)	Observed Frequency	$p(x) = {}^{7}C_{x} (0.4833)^{x} (0.5167)^{7-x}$	Expected or Theoretical Frequency , f(x)=N. p(x)
0	7	$p(0) = (0.5167)^{7} = 0.0098$	$f(0) = 128.p(0) = 1.25$
1	6	$p(1) = {}^{7}C_{1} (0.4833)^{1} (0.5167)^{7-1} = 0.0643$	$f(1) = 128.p(1) = 8.2304$
2	10	$p(2) = {}^{7}C_{1} (0.4833)^{2} (0.5167)^{7-2} = 0.1806$	$f(2) = 128.p(2) = 23.11$
3	35	$p(3) = 0.2816$	$f(3) = 128.p(3) = 36.04$
4	30	$p(4) = 0.2634$	$f(4) = 128.p(4) = 33.71$
5	23	$p(5) = 0.1478$	$f(5) = 128.p(5) = 18.91$
6	7	$p(6) = 0.0461$	$f(6) = 128.p(6) = 5.89$
7	1	$p(7) = 0.0061$	$f(7) = 128.p(7) = 0.78$

Exercise 3:

1. Five dice were thrown together 96 times. The number of times 4, 5, 6 was actually thrown is given below. Calculate the expected frequency.

x	0	1	2	3	4	5
$f(x)$	1	10	24	35	18	8

2. Fit a Binomial Distribution to the following data:

x	0	1	2	3	4	5	6
$f(x)$	6	28	56	60	36	12	2

Poisson Distribution

Poisson distribution is a limiting case of the 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 for each trial is indefinitely small, i.e., $p \to 0$.

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

Definition:

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

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

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

Note: It should be noted that

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

Hence equation (1) is a probability function.

Examples of Poisson distribution:

1. The number of defective electric bulbs manufactured by a reputed company.
2. The number of telephone calls per minute at a switch board.
3. The number of cars passing a certain point in one minute.
4. The number of printing mistakes per page in a large text.
5. The number of particles emitted by a radio-active substance.
6. The number of persons born blind per year in a large city.

Constants of Poisson distribution

Mean of the Poisson distribution:

Mean of X, $\mu = E(X) = \sum_{x=0}^{n} x\, p(x) = \sum_{x=0}^{\infty} x.\frac{e^{-\lambda}\lambda^{x}}{x!} = e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x}}{(x-1)!}$ $\qquad [\because x! = x(x-1)!]$

$$= e^{-\lambda}\left[\lambda + \frac{\lambda^{2}}{1!} + \frac{\lambda^{3}}{2!} + ...\right] = e^{-\lambda}\lambda\left[1 + \frac{\lambda}{1!} + \frac{\lambda^{2}}{2!} + ...\right] = e^{-\lambda}\lambda e^{\lambda} = \lambda$$

Thus the parameter λ is the Mean of the Poisson distribution.

Variance of Poisson Distribution:

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

$$= \sum_{x=0}^{n} x^2\, p(x) - \mu^2 = \sum_{x=0}^{n} \left(x(x-1)+x\right) p(x) - \mu^2$$

$$= \sum_{x=0}^{n} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{n} xp(x) - \lambda^2$$

$$= e^{-\lambda}\sum_{x=2}^{n}\frac{\lambda^x}{(x-2)!} + \mu - \mu^2 = e^{-\lambda}\lambda^2\sum_{x=2}^{n}\frac{\lambda^{x-2}}{(x-2)!} + \lambda - \lambda^2$$

$$= e^{-\lambda}\lambda^2 e^{\lambda} + \lambda - \lambda^2 = \lambda$$

$\therefore$ Variance of the Binomial distribution $= \lambda$

Hence the Standard Deviation of Binomial distribution $= \sqrt{\lambda}$

Example 1: If the probability that an individual suffers a bad reaction from a certain injection is 0.001, determine the probability that out of 2000 individuals (i) exactly 3 (ii) more than 2 individuals (iii) none (iv) more than one individual suffer a bad reaction.

Solution: Given p=0.001 and n=2000

$\therefore$ Mean, $\lambda = np = 2000\ (0.001) = 2$

$$\therefore \quad P(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-2}2^x}{x!}$$

(i) $$P(3) = \frac{e^{-2}2^3}{3!} = \frac{8}{6(2.718)^2} = 0.1804$$

(ii) $$P(\text{more than } 2) = P(X \ge 2) = P(3) + P(4) + \ldots + P(2000)$$

$$= 1 - \left[P(0) + P(1) + P(2)\right] = 1 - \left[\frac{1}{e^2} + \frac{2}{e^2} + \frac{4}{2e^2}\right] = 1 - \frac{5}{e^2} = 1 - 0.67667 = 0.3233$$

(iii) $$P(\text{none}) = P(0) = \frac{e^{-2}2^0}{0!} = \frac{1}{e^2} = 0.1353$$

(iv) $$P(\text{more than one}) = P(X > 1) = P(2) + P(3) + \ldots + P(2000) = 1 - \left[P(0) + P(1)\right]$$

$$= 1 - \left(\frac{1}{e^2} + \frac{2}{e^2}\right) = 1 - \frac{3}{e^2} = 1 - 0.406 = 0.594$$

Example 2: A car-hire firm has two cars which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with mean 1.5. Calculate the proportion of days (i) on which there is no demand (ii) on which demand is refused.

Solution: Given mean, $\lambda = 1.5$

We have $P(x) = \dfrac{e^{-\lambda}\lambda^{x}}{x!}$

(i) P(no demand) = $P(0) = \dfrac{e^{-15}(1.5)^{0}}{0!} = e^{-1.5} = 0.2231$

Note: Number of days in a year there is no demand of car = 365(0.2231) = 81 days

(ii) Some demand is refused if the number of demands is more than two i.e., $X > 2$

P(demand refused) = $P(X>2) = 1 - [P(0)+P(1)+P(2)]$

$$= 1 - \left[e^{-1.5} + \frac{e^{-1.5}(1.5)}{1!} + \frac{e^{-1.5}(1.5)^{2}}{2!}\right] = 1 - e^{-1.5}[1+1.5+1.125]$$

$$= 1 - 3.625\left(e^{-1.5}\right) = 1 - 0.8088 = 0.1913$$

Note: Number of days in a year when some demand is refused

= 365 X 0.1913 = 69.82 = 70 days

Example 3: If a Poisson distribution is such that $P(X=1)\dfrac{3}{2} = P(X=3)$, find

(i) $P(X \geq 1)$ (ii) $P(X \leq 3)$ (iii) $P(2 \leq X \leq 5)$

Solution: Given $\dfrac{3}{2}P(X=1) = P(X=3)$

i.e., $\dfrac{3}{2}.\dfrac{e^{-\lambda}.\lambda^{1}}{1!} = \dfrac{e^{-\lambda}.\lambda^{3}}{3!}$ *i.e.,* $\dfrac{3\lambda}{2} = \dfrac{\lambda^{3}}{6}$

$\lambda^{3} - 9\lambda = 0$ or $\lambda\left(\lambda^{2}-9\right) = 0$ or $\lambda(\lambda-3)(\lambda+3) = 0$

$\therefore \lambda = 0, 3, -3$ $\Rightarrow \lambda = 3$ $(\because \lambda > 0)$

Hence $P(X=x) = p(x) = \dfrac{e^{-3}3^{x}}{x!}$

(i) $P(X \geq 1) = 1 - P(x<1) = 1 - P(X=0) = 1 - \dfrac{e^{-3}.3^{0}}{0!} = 1 - e^{-3} = 0.950213$

(ii) $P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$

$$= e^{-3}\left[\frac{3^0}{0!}+\frac{3}{1!}+\frac{3^2}{2!}+\frac{3^3}{3!}\right] = e^{-3}\left(1+3+\frac{9}{2}+\frac{9}{2}\right) = 13e^{-3} = 0.6472318$$

(iii) $P(2 \le X \le 5) = P(X=2)+P(X=3)+P(X=4)+P(X=5)$

$$= e^{-3}\left[\frac{3^2}{2!}+\frac{3^3}{3!}+\frac{3^4}{4!}+\frac{3^5}{5!}\right] = 9e^{-3}\left(\frac{1}{2}+\frac{1}{2}+\frac{3}{8}+\frac{9}{40}\right) = 9e^{-3}\left(1.6\right) = 0.7169337$$

Exercise 4:

1. If X is Poisson variate such that $P\left(X=0\right)=P\left(X=2\right)+3P\left(X=4\right)$,
   find (i)the mean of X (ii) $P(X \le 2)$ [Ans: (i) 1.21, (ii) 0. 87731]
2. A manufacturer of cotter pins knows that 5% of his product is defective. If he sells cotter pins in boxes of 100 and generates that not more than 10 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality? [Ans: 0.0136]
3. An insurance company insures 4,000 people against loss of both eyes in a car accident. Based on previous data, the rates were computed on the assumption that on the average 10 persons in 1,00,000 will have car accident each year that result in this type of injury. What is the probability that more than 3 of the insured will collect on their policy in a given year? [Ans: 0.0008]
4. In a frequency distribution, frequency corresponding to 3 successes is $2/3$ times frequency corresponding to 4 successes. Find the mean and standard deviation of the distribution. [Ans: 6]

Note: P (X successes during given time interval by Poisson distribution is) = $\lambda = np = \alpha T$

α is average no:of successes per unit time

Example 4: The average number of phone calls / minute coming into a switch board between 2p.m and 4p.m is 2.5. Determine the probability that during one particular minute there will be
(i) 4 or fewer (ii) more than 6 calls.

Solution: Let X be the number of phone calls / minute coming into a switch board.

Given mean, $\lambda = 2.5$

Now the Poisson distribution is $p(x) = \dfrac{e^{-\lambda}.\lambda^x}{x!} = \dfrac{e^{-2.5}\left(2.5\right)^x}{x!}$

(i) $P(X \le 4) = P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)$

$$= e^{-2.5}\left[\frac{\left(2.5\right)^0}{0!}+\frac{\left(2.5\right)^1}{1!}+\frac{\left(2.5\right)^2}{2!}+\frac{\left(2.5\right)^3}{3!}+\frac{\left(2.5\right)^4}{4!}\right] = 0.8912$$

(ii) $$P(X>6)=1-P(X\le 6)=1-e^{-2.5}\left[\frac{(2.5)^0}{0!}+\frac{(2.5)^1}{1!}+\ldots+\frac{(2.5)^6}{6!}\right]=0.01416$$

Exercise 5:

1. The average number of phone calls arriving at a telephone exchange is 30 per hour. What is the probability that (i) no calls arrive in a 3 minute period
(ii) more than 5 calls arrive in 5 minute period [Ans: (i) 0.2231, (ii) 0.642]

2. A manufacture accepts the work submitted by his typist only when there is no mistake in the work. The typist have to type on an average 20 letters per day of about 200 words each. Find the chance of her making a mistake
(i) if less than 1% of the letters submitted by her are rejected, [Ans: 0.0000506]
(ii) if on 90% days all the letters submitted by her are accepted . [Ans: 0.0000263]

3. In a book of 520 pages, 390 typo-graphical errors occur. Assuming Poisson law for the number of errors per page, find the probability that a random sample of 5 pages will contain no error. [Ans: 0.0235]

4. Average number of accidents on any day on a national highway is 1.8. Determine the probability that the number of accidents is (i) at least one (ii) at most one [Ans: (i) 0.8347, (ii) 0.4628]

5. If a bank receives on the average 6 bad cheques per day. What is the probability that it will receive (i) 4 bad cheques on any given day (ii) 10 bad cheques over any consecutive days.

Fitting of Poisson distribution:

Recurrence Relation for the probabilities of Poisson distribution:

We have $\frac{p(x+1)}{p(x)}=\frac{e^{-\lambda}\lambda^{x+1}}{(x+1)!}\cdot\frac{x!}{e^{-\lambda}\lambda^{x}}=\frac{\lambda}{x+1}$ (on simplification)

$\therefore\ p(x+1)=\left(\frac{\lambda}{x+1}\right).p(x)$ is recurrence formula for the probabilities of Poisson Distribution.

Example 1: Fit a Poisson distribution for the following data and calculate the expected frequencies

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

Solution: Here N=total frequency = $\sum f_1 = 109+65+22+3+1=200$

$$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i}=\frac{0+65+44+9+4}{200}=\frac{122}{200}=0.61 \quad \Rightarrow \lambda=0.61$$

$\therefore$ Mean of Poisson distribution, $\lambda=0.61$

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

x	Observed Frequency	Probability , $p(x)$	Expected or Theoretical Frequency $f(x) = N.p(x) = 200.p(x)$
0	109	$e^{-0.61}$	$f(0) = 200.p(0) = 108.67 \approx 109$
1	65	$e^{-0.61}(0.61)$	$f(1) = 200.p(1) = 66.29 \approx 66$
2	22	$e^{-0.61}.\frac{(0.61)^2}{2!}$	$f(2) = 200.p(2) = 20.22 \approx 20$
3	3	$e^{-0.61}.\frac{(0.61)^3}{3!}$	$f(3) = 200.p(3) = 4.11 \approx 4$
4	1	$e^{-0.61}.\frac{(0.61)^4}{4!}$	$f(4) = 200.p(4) = 0.63 \approx 1$

Example 2: The distribution of typing mistakes committed by a typist is given below. Assuming the distribution to be Poisson, find the expected frequencies

Number of mistakes per page, x	0	1	2	3	4	5
Number of pages, $f(x)$	42	33	14	6	4	1

Solution: $N = \sum f_i = 100$

Mean of the distribution, $\lambda = \frac{\sum f_i x_i}{\sum f_i} = \frac{0+33+28+18+16+5}{100} = \frac{100}{100} = 1$

By Recurrence relation, $\therefore\ p(x+1) = \left(\frac{\lambda}{x+1}\right).p(x) = \left(\frac{1}{x+1}\right).p(x)$

$$P(0) = e^{-1} = 0.3678$$

(x)	Observed Frequency	Probability , $p(x)$	Expected or Theoretical Frequency $f(x) = N.p(x) = 100.p(x)$
0	42	$e^{-1} = 0.3678$	$f(0) = 100.p(0) = 36.79 \approx 37$
1	33	$p(1) = \frac{1}{0+1}p(0) = 0.3678$	$f(1) = 100.p(1) = 36.79 \approx 37$
2	14	$p(2) = \frac{1}{1+1}p(1) = 0.1839$	$f(2) = 100.p(2) = 18.39 \approx 18$
3	6	0.0613	$f(3) = 100.p(3) = 6.13 \approx 6$
4	4	0.0153	$f(4) = 100.p(4) = 1.53 \approx 2$
5	1	0.0031	$f(5) = 100.p(5) = 0.31 \approx 0$

Exercise 6:

1. After correcting 50 pages of the proof of a book, the proof reader finds that there are, on the average, 2 errors per 5 pages. How many pages would one expect to find with 0, 1, 2, 3 and 4 errors, in 1,000 pages of the first print of the book?

Uniform Distribution

This Section introduces the continuous probability distribution which features a continuous random variable X with probability density function $f(x)$ which assumes a constant value over a finite interval.

Definition:

A random variable X is said to have a continuous uniform distribution over an interval (a,b) if its probability density function is constant = k (say), over the entire range of X , i.e.,

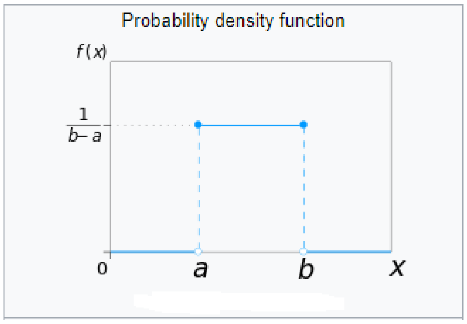
$$f(x)=\begin{cases} k; a<x<b \\ 0, \text{otherwise} \end{cases}$$

Since the total probability is always unity, $\int_a^b f(x)dx=1 \quad \Rightarrow k\int_a^b dx=1 \quad \Rightarrow k=\frac{1}{b-a}$

$\therefore$ The function $f(x)$ is defined by $f(x)=\begin{cases} \frac{1}{b-a}; a<x<b \\ 0, \text{otherwise} \end{cases}$

Note:

7. a and $b, (a<b)$ are two parameters of the uniform distribution on (a,b).
8. The distribution is also known as rectangular distribution, since the curve $y=f(x)$ describes a rectangle over the x-axis and between the ordinates at $x=a$ and $x=b$. The graph of uniform distribution is



9. The distribution function $F(x)$ is given by

$$F(x)=\begin{cases} 0, -\infty<x<a \\ \frac{x-a}{b-a}, a\le x\le b \\ 1, b<x<\infty \end{cases}$$

Constants of Uniform distribution

Mean of the Uniform distribution:

Mean of X, $\mu = E(X) = \int_a^b x\, f(x)dx = \int_a^b x \frac{1}{b-a}dx = \frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right) = \frac{b+a}{2}$

Variance of Uniform Distribution:

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

$$= \int_a^b x^{\,2} f(x)dx - \mu^2 = \int_a^b x^{\,2}\frac{1}{b-a}dx - \left(\frac{b+a}{2}\right)^2$$

$$= \frac{1}{b-a}\left(\frac{b^3-a^3}{3}\right) - \left(\frac{b+a}{2}\right)^2 = \frac{1}{b-a}\left(\frac{(b-a)\left(b^2+ab+a^2\right)}{3}\right) - \left(\frac{b^2+2ab+a^2}{4}\right)$$

$$= \frac{(b-a)^{\,2}}{12}$$

Example 1: If X is uniformly distributed with mean 1 and variance $^4\!/_3$. Find $P(X<0)$.

Solution: Given mean = 1 and variance = $^4\!/_3$

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

$Var(X) = \frac{(b-a)^2}{12} = \frac{4}{3} \quad \Rightarrow (b-a)^2 = 16 \;\Rightarrow b-a = \pm 4$

Solving, $a=-1$ and $b=3\,(\because a<b)$

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

$P(X<0) = \int_{-1}^{0} \frac{1}{4}dx = \frac{1}{4}$

Example 2: The thickness of a protective coating applied to a conductor designed to work in corrosive conditions follows a uniform distribution over the interval $[20,40]$ microns. Find the mean, standard deviation and cumulative distribution function of the protective coating. Find also the probability that the coating is less than 35 microns thick.

Solution: Over the interval $[20,40]$ the probability density function (p.d.f) $f(x)$ is given by

$$f(x)=\begin{cases}\dfrac{1}{20}=0.05; 20<x<40\\ 0, \text{otherwise}\end{cases}$$

$$\therefore \text{ Mean} = \frac{b+a}{2}=10\mu m$$

$$Var(X)=\frac{(b-a)^2}{12}=\frac{400}{12} \Rightarrow (b-a)^2=\frac{100}{3}$$

$$\Rightarrow S.D = 5.77\mu m$$

The cumulative distribution function is $F(x)=\int_{-\infty}^{x} f(x)dx=\begin{cases}0, x<20\\ \dfrac{x-20}{20}, 20\le x\le 40\\ 1, x\ge 40\end{cases}$

The probability that the coating is less than 35 microns thick is $F(X<35)=\dfrac{35-40}{20}=0.75$

Exercise 7:

1. Subway trains on a certain line run every half hour between mid-night and six in the morning. What is the probability that a man entering the station at a random time during this period will have to wait at least twenty minutes? [Ans: 1/3]

2. A random variable X has a uniform distribution over $(-3,3)$, Compute

(i) $P(X=2)$, $P(X<2)$, $P(|X|<2)$ and $P(|X-2|<2)$ [Ans:]

3. The current (in mA) measured in a piece of copper wire is known to follow a uniform distribution over the interval $[0,25]$. Write the probability density function (p.d.f) $f(x)$ of the random variable X representing the current. Calculate the mean and variance of the distribution.

[Ans: 12.5mA , 52.08 mA²]

Normal Distribution

Normal distribution is a continuous distribution of fundamental importance. Any quantity whose variation depends on random causes is distributed according to the normal law. Its importance lies in the fact that a large number of distributions approximate to normal distribution

Definition:

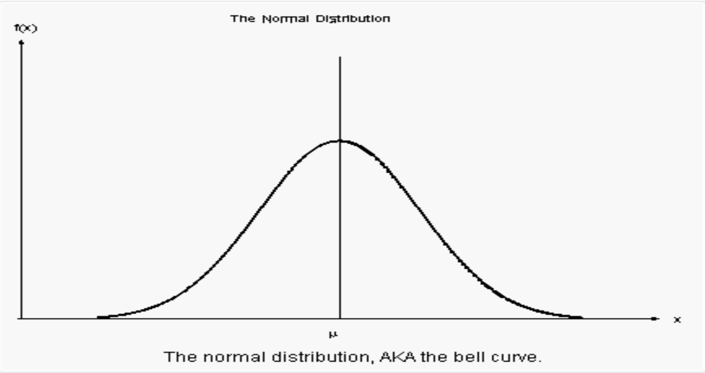
A random variable X is said to have a Normal distribution, if its density function or probability distribution is given by

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

Where μ =Mean and σ = standard deviation are two parameters of the normal distribution.

The random variable X is then said to be a normal random variable or normal variate. The curve representing the Normal Distribution (1) is called the normal curve and the total area bounded by the curve and the x-axis is one

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



The Normal Distribution

The normal distribution, AKA the bell curve.

The area under the curve between the ordinates $x=a$ and $x=b$, where a<b, represents the probability that x lies between a and b. i.e., $P(a<x<b)$

Thus $P(a<x<b)$ = area under the normal curve between the vertical lines $x=a$ and $x=b$, which is $\int_{a}^{b} f(x)dx$

Note: A random variable X with mean μ and variance σ^2 and following the probability law (1) is expressed by $X \sim N\left(\mu,\sigma^2\right)$

Applications of Normal distribution:

1. Calculations of errors made by chance in experimental measurements.
2. Computation of hit probability of a shot.
3. Statistical inference in almost every branch of science.

Constants of Normal distribution

Mean of the Normal distribution:

Consider the normal distribution with b,σ as the parameters. Then $f(x;b,\sigma)=\dfrac{e^{-\frac{(x-b)^2}{2\sigma^2}}}{\sigma.\sqrt{2\pi}}$

Mean of X, $\mu=\int_{-\infty}^{\infty} xf(x)dx=\frac{1}{\sigma.\sqrt{2\pi}}\int_{-\infty}^{\infty} x.e^{-\frac{1}{2}\left(\frac{x-b}{\sigma}\right)^2}dx$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma z+b)e^{-\frac{z^2}{2}}dx \qquad \left[\text{putting } z=\frac{x-b}{\sigma} \text{ so that } dz=\frac{dx}{\sigma}\right]$$

$$=\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}ze^{-\frac{z^2}{2}}dx+\frac{b}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{z^2}{2}}dx$$

$$=0+\frac{b}{\sqrt{2\pi}}.2\int_{0}^{\infty}e^{-\frac{z^2}{2}}dz \quad \left[\because ze^{-\frac{z^2}{2}}\text{is odd function and } e^{-\frac{z^2}{2}}\text{is even function}\right]$$

$$=\frac{2b}{\sqrt{2\pi}}.\frac{\sqrt{\pi}}{\sqrt{2}} \quad \left[\because\int_{0}^{\infty}e^{-\frac{x^2}{x}}dx=\sqrt{\frac{\pi}{2}}\right]=b$$

$\therefore$ Mean, $\mu=b$

Variance of Normal Distribution:

Variance $=V(x)=E\left(X^2\right)-\left[E(X)\right]^2=\int_{-\infty}^{\infty}(X-\mu)^2 f(x)dx$

$$=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-b)^2e^{-\frac{1}{2}\left(\frac{x-b}{\sigma}\right)^2}dx \qquad =\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma z)^2e^{-\frac{z^2}{2}}dz \qquad \left[\text{Putting } z=\frac{x-b}{\sigma} \text{ so that } dz=\frac{dx}{\sigma}\right]$$

$$=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}z^2e^{-\frac{z^2}{2}}dz \qquad =\frac{\sigma^2}{\sqrt{2\pi}}.2\int_{0}^{\infty}z^2e^{-\frac{z^2}{2}}dz \qquad [\because \text{ Integrand is even function}]$$

$$=\frac{2\sigma^2}{\sqrt{2\pi}}\int_{0}^{\infty}2te^{-t}.\frac{dt}{\sqrt{2t}} \qquad \left[\text{Putting } \frac{z^2}{2}\text{=t so that } dz=\frac{dt}{\sqrt{2t}}\right]$$

$$=\frac{2\sigma^2}{\sqrt{\pi}}\int_{0}^{\infty}e^{-t}\sqrt{t}dt=\frac{2\sigma^2}{\sqrt{\pi}}\int_{0}^{\infty}e^{-t}t^{\frac{3}{2}-1}dt=\frac{2\sigma^2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) \quad =\frac{2\sigma^2}{\sqrt{\pi}}.\frac{1}{2}.\tau\left(\frac{1}{2}\right) \quad =\frac{2\sigma^2}{\sqrt{\pi}}.\frac{\sqrt{\pi}}{2} \quad =\sigma^2$$

Hence Variance = σ^2

Thus the Standard deviation of the Normal Distribution is σ

Mode of Normal Distribution:

Mode is the value of x for which f(x) is maximum. That is, mode is the solution of $f^{'}(x)=0 \quad and \quad f^{''}(x)<0$

By definition, we have $f(x)=\frac{1}{\sigma.\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Differentiating w.r.t x , we get

$$f^{'}(x)=\frac{1}{\sigma.\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\left[-\left(\frac{x-\mu}{\sigma}\right)\right]=-\frac{x-\mu}{\sigma}\frac{1}{6}f(x)$$

Now $f^{'}\left(x\right)=0 \Rightarrow x-\mu=0$ i.e., $x=\mu$

$$f^{"}(x)=-\frac{1}{\sigma^2}\left[\left(x-\mu\right).f^{1}(x)+f(x)\right]=-\frac{1}{\sigma^2}\left[\left(x-\mu\right)-\frac{\left(x-\mu\right)}{\sigma^2}.f(x)+f(x)\right]=\frac{-f(x)}{\sigma^2}\left[1-\frac{\left(x-\mu\right)^2}{\sigma^2}\right]$$

At the point $x=\mu$, we have $f^{"}(x)=-\left[\frac{f(x)}{\sigma^2}\right]_{x=\mu}=-\frac{1}{\sigma^2.\sqrt{2\pi}.\sigma}<0$

Hence $x=\mu$ is the mode of the Normal Distribution

Median of Normal distribution:

If M is the median of the normal distribution, we have $\int\limits_{-\infty}^{M} f(x)dx=\frac{1}{2}$

i.e., $$\frac{1}{\sigma.\sqrt{2\pi}}\int\limits_{-\infty}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=\frac{1}{2}$$

i.e., $$\frac{1}{\sigma.\sqrt{2\pi}}\int\limits_{-\infty}^{\mu} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx+\frac{1}{\sigma.\sqrt{2\pi}}\int\limits_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=\frac{1}{2} \qquad \text{... (1)}$$

Consider $\frac{1}{\sigma.\sqrt{2\pi}}\int\limits_{-\infty}^{\mu} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx$ Put $\frac{x-\mu}{\sigma}=z$. Then $dx=\sigma dz$

$$\therefore \frac{1}{\sigma.\sqrt{2\pi}}\int\limits_{-\infty}^{\mu} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=\frac{1}{\sigma.\sqrt{2\pi}}\int\limits_{-\infty}^{0} e^{\frac{x^2}{2}}.\sigma dz=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{0} e^{\frac{x^2}{2}}dz$$

$$=\frac{1}{\sqrt{2\pi}}\int\limits_{0}^{\infty} e^{-\frac{z^2}{2}}dz \text{ (by symmetry) }=\frac{1}{\sqrt{2\pi}}.\sqrt{\frac{\pi}{2}}=\frac{1}{2} \qquad \text{... (2)}$$

$\therefore$ From (1) and (2), we have $\frac{1}{2}+\frac{1}{\sigma.\sqrt{2\pi}}\int\limits_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=\frac{1}{2}$

$$\Rightarrow \frac{1}{\sigma.\sqrt{2\pi}}\int\limits_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=0 \qquad \Rightarrow \int\limits_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=0$$

$$\Rightarrow \quad \mu=M \quad \left[\because \text{ if } \int\limits_{a}^{b} f(x)dx=0 \text{ then } a=b, \text{ where } f(x)>0\right]$$

Hence for the normal distribution, Mean=Median

Note: From above, we notice that for the Normal distribution mean, median and mode coincide. i.e., Mean=Median=Mode . Hence the distribution is symmetrical.

Chief Characteristics of the Normal Distribution:

1. The graph of the Normal distribution $y = f(x)$ in the xy-plane is known as the normal curve.
2. The curve is bell shaped and symmetrical about the line $x = \mu$ and the two tails on the right and the left sides of the mean (μ) extends to infinity.
3. Area under the normal curve represents the total population.
4. Mean, median and mode of the distribution coincide. So normal curve is unimodal (has only one maximum point).
5. X-axis is an asymptote to the curve.
6. Linear combination of independent normal varieties is also a normal variate.
7. The points of inflexion of the curve are at $x = \mu \pm \sigma$
8. The probability that the normal variate X with mean μ and standard deviation σ lies between x_1 and x_2 is given by

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

Since (1) depends on the two parameters μ and σ, we get different normal curves for different values of μ and σ and it is an impracticable task to plot all such normal curves. Instead, by putting $z = \frac{x-\mu}{\sigma}$ the R.H.S of equation (1) becomes independent of the two parameters μ and σ. Here z is known as the standard variable.

9. Area under the normal curve is distributed as follows:
   (i) Area of normal curve between $\mu - \sigma$ and $\mu + \sigma$ is 68.27%
   i.e, $P(\mu - \sigma < X < \mu + \sigma) = 0.6826$
   (ii) Area of normal curve between $\mu - 2\sigma$ and $\mu + 2\sigma$ is 95.43%.
   (iii) Area of normal curve between $\mu - 3\sigma$ and $\mu + 3\sigma$ is 99.73%.

Handout # 25

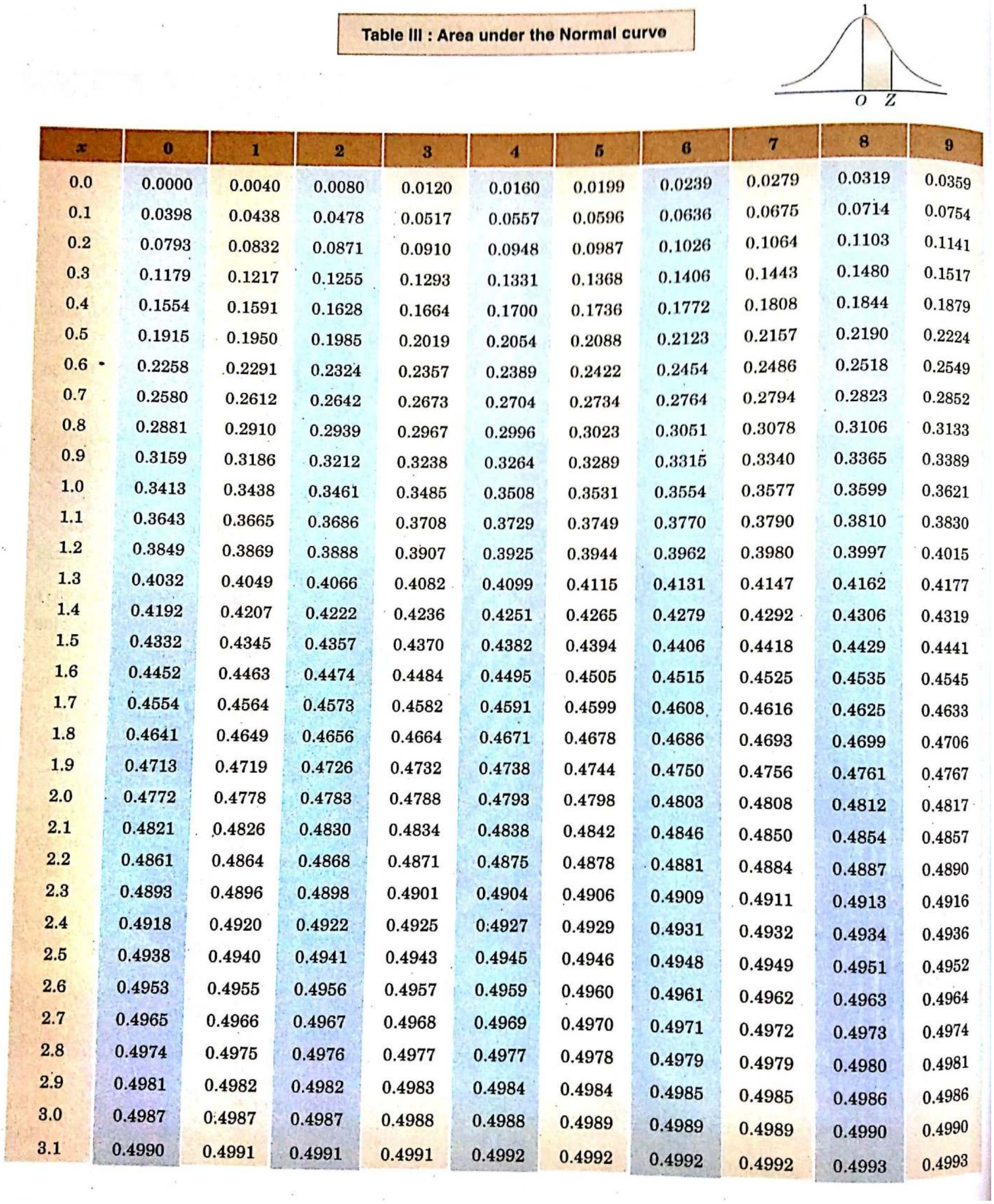


Table III : Area under the Normal curve

x	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2258	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2996	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993

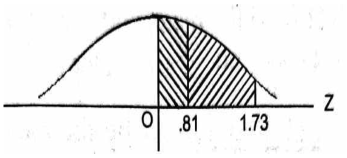
Examples on Normal Distribution:

Example 1: For a normally distributed variate with mean 1 and standard deviation 3, find the probabilities that (i) $3.43 \le x \le 6.19$ (ii) $-1.43 \le x \le 6.19$

Solution: Given $\mu = 1$ and $\sigma = 3$

(i) When $x = 3.43$, $z = \dfrac{x-\mu}{\sigma} = \dfrac{3.43-1}{3} = \dfrac{2.43}{3} = 0.81 = z_1$ (say)

When $x = 6.19$, $z = \dfrac{x-\mu}{\sigma} = \dfrac{6.19-1}{3} = \dfrac{5.19}{3} = 1.73 = z_2$ (say)



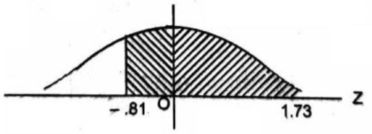
$$\therefore P(3.43 \le x \le 6.19) = P(0.81 \le x \le 1.73) = \left|A(z_2) - A(z_1)\right|$$

$$= \left|A(1.73) - A(0.81)\right| = 0.4852 - 2910 \text{ (from tables)}$$

= 0.1672 (Cross hatched area in the figure)

(ii) When $x = -1.43$, $z = \dfrac{x-\mu}{\sigma} = \dfrac{-1.43-1}{3} = -0.81 = z_1$ (say)

When $x = 6.19$, $z = \dfrac{x-\mu}{\sigma} = \dfrac{6.19-1}{3} = 1.73 = z_2$ (say)



$$\therefore P(-1.43 \le x \le 6.19) = P(-0.81 \le x \le 1.73) = A(z_2) + A(z_1)$$

$$= A(1.73) + A(0.81) \quad [\because A(-z) = A(z)]$$

= 0.4582 + 0.2910 = 0.7492 = shaded area in the figure.

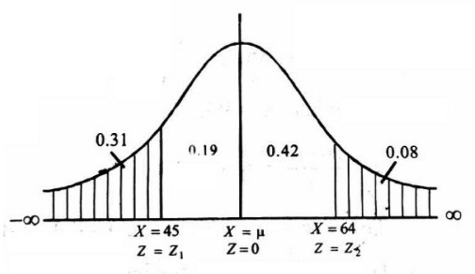
Example 2: In a normal distribution 31% of the items are under 45 and 8% are over 64. Find the mean and variance of the distribution.

Solution: Let X be the continuous random variable. Let μ be the mean and σ the standard deviation.

Given $P(X < 45) = 0.31$ *and* $P(X > 64) = 0.08$

When $X = 45$, *let* $z = z_1$ so that $-z_1 = \dfrac{45-\mu}{\sigma}$ .... (1)

When X=64, *let* $z = z_2$ so that $z_2 = \dfrac{64-\mu}{\sigma}$ ... (2)



Hence $P(-z_1 \le z \le 0) = P(0 < z < z_1) = 0.5 - 0.31 = 0.19$

The corresponding z_1 value for the area 0.19 is $z_1 = 0.5$ (from table) ... (3)

And $P(0 < z < z_2) = 0.5 - 0.08 = 0.42$

The corresponding z_2 value for the area 0.42 is $z_1 = 1.4$ (from table) ... (4)

From (1) and (3), we have $\dfrac{45-\mu}{\sigma} = -0.5 \Rightarrow 45 - \mu = -0.5\sigma$ ...(5)

From (2) and (4), we have $\dfrac{64-\mu}{\sigma} = 1.4 \Rightarrow 64 - \mu = 1.4\sigma$ ...(6)

Solving (5) and (6) $\mu = 50$ and $\sigma = 10$

Example 3: The marks obtained in mathematics by 1000 students is normally distributed with mean 78% and standard deviation 11%. Determine

(i) How many students got marks above 90%

(ii) What was the highest mark obtained by the lowest 10% of the students

(iii) Within what limits did the middle of 90% of the students lie.

Solution: Given mean $\mu = 0.78$, and S.D $\sigma = 0.11$

(i) When $x = 0.9,\ z = \dfrac{x-\mu}{\sigma} = \dfrac{0.9-0.78}{0.11} = 1.09 = z_1$ (say)

Hence the number of students with marks more than 90%= 0.1379 X 1000= 137.9 138

(ii) The 0.1 area to the left of z corresponds to the lowest 10% of the students

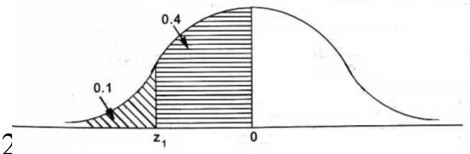


Fig.

From Figure, 0.4 = 0.5 – 0.1 = 0.5 – Area from 0 to z_1

$\therefore \quad z_1 = -1.28$ (from tables)

Thus $-1.28 = \dfrac{x-\mu}{\sigma} = \dfrac{x-0.78}{0.11} \Rightarrow x = 0.78 - 1.28(0.11) = 0.6392$

Hence the highest mark obtained by the lowest 10% of students= 0.6392 X 1000 64%

(iii) Middle 90% correspond to 0.9 area, leaving 0.05 area on both sides.

Then the corresponding z's are ± 1.64

[since if the area from 0 to z is 0.45, then z = 1.64]

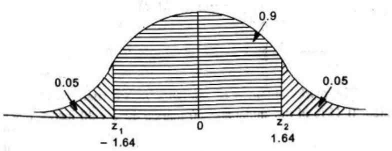


Fig.

$\therefore -1.64 = z_1 = \dfrac{x_1-\mu}{\sigma} = \dfrac{x_1-0.78}{0.11}$ and $1.64 = z_2 = \dfrac{x_2-\mu}{\sigma} = \dfrac{x_2-0.78}{0.11}$

Solving, $x_1 = 0.5996\, or\, 59.96\%$ and $x_2 = 0.9604\, or\, 96.04\%$

Thus the middle 90% have marks in between 60 to 96.

Example 4: In a sample of 1000 cases, the mean of a certain test is 14 and standard deviation is 2.5. Assuming the distribution to the normal, find

(i) How many students score between 12 and 15?

(ii) How many score above 18?

(iii) How many score below 18?

Solution: Let be the mean μ and σ the standard deviation of the normal distribution.

Then we are given $\mu = 14$ and $\sigma = 2.5$

Let the variable X denote the score in a test.

(i) When $X = 12, z = \dfrac{x-\mu}{\sigma} = \dfrac{12-14}{2.5} = -0.8 = z_1$ (say)

$\therefore \quad P(12 < X < 15) = P(-0.8 < z < 0.4)$

$= A(z_2) + A(z_1) = A(0.4) + A(-0.8) = A(0.4) + A(0.8)$ (due to symmetry)

=0.1554+0.2881=0.4435

$\therefore$ Number of students score between 12 and 15 =1000 X 0.4435 = 443 (approximately)

(ii) When $X = 18, z = \dfrac{x-\mu}{\sigma} = \dfrac{18-14}{2.5} = 1.6$

$\therefore$ $P(X > 18) = P(z > 1.6)$ = 0.5 - A(1.6) = 0.5 - 0.4452= 0.0548

$\therefore$ Number of students score above 18=1000 X 0.0548 = 54.8 = 55 (approximately)

(iii) When score below 18

$P(X < 18) = P(z < 1.6)$ =0.5 + A(1.6) = 0.5 + 0.4452 = 0.9452

Aliter: $P(X < 18) = 1 - P(X > 18) = 1 - 0.0548 = 0.9452$

$\therefore$ Number of students score below 18 = 1000

Exercise 8:

1. If X is a normal variate, find the area A

(i) to the left of $z = -1.78$ (ii) to the right of $z = -1.45$ (iii) to the left of $z = 0.56$

(iv) corresponding to $z \geq 2.16$ (v) corresponding to $-0.8 \leq z \leq 1.53$

(vi) to the left of $z = -2.52$ and to the right of $z = 1.83$

2. X is a normal variate with mean 30 and S.D 5. Find the probabilities that

(i) $26 \leq X \leq 40$ (ii) $X \geq 45$ and (iii) $|X - 30| > 5$ [Ans:0.7653, 0.00135, 0.3174]

3. The mean yield for one-acre plot is 662 kilos with S.D 32 kilos. Assuming normal distribution, how many one-acre plots in a batch of 1,000 plots would you expect to have yield

(i) over 700 kilos, (ii) below 650 kilos, and (iii) what is the lowest yield of the best 100 plots?

[Ans: 117, 352, 702.96 kilos]

4. In a distribution exactly normal, 10.03% of the items are under 25 kilogram weight and 89.97% of the items are under 70 kilogram weight. What are the mean and standard deviation of the distribution? [Ans: 47.5, 17.578 kilogram]

5. In a test of 2000 electrical bulbs, it was found that the life of a particular make was normally distributed with an average life of 2040 hours and S.D of 40 hrs. Estimate the number of bulbs likely to burn for

(i) more than 2140 hours (ii) less than 1950 hours

(iii) more than 1920 hours and less than 2160 hours [Ans: 12, 134, 1994]

6. Suppose the weights of 200 students are normally distributed with mean 140 pounds and S.D 10 pounds. Find the number of students weights are

(i) between 138 and 148 pounds (ii) more than 152 pounds [Ans: 294, 92]

Computer Oriented Statistical Methods (A8005)
#3.1 Correlation and Regression

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

The correlation expresses the relationship or interdependence of two sets of variables upon each other. One variable may be called the subject (independent) and other relative (dependent).

Types of Correlation:

Correlation is classified into many types.

1. Positive and negative
2. Simple and multiple
3. Partial and total
4. Linear and non-linear

1. Positive and Negative Correlation:

Positive and Negative correlation depend upon the direction of change of the variables. If two variables tend to move together in the same direction i.e an increase in the value of one variable is accompanied by an increase in the value of the other variable; or a decrease in the value of one variable is accompanied by a decrease in the value of the other variable, then the correlation is called positive or direct correlation. Height and weight, rainfall and yield of crops, price and supply are examples of positive correlation.

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

2. Simple and multiple correlations:

When we study two variables, the relationship is described as simple correlation example are quantity of money and price level, demand and price etc. But in multiple correlation we study more than two variables simultaneously; example is the relationship of price, demand and supply of commodity.

3. Partial and total correlation:

The study of two variables excluding some other variables is called partial correlation. For example, we study price and demand, eliminating the supply side. In total correlation, all fact are taken into account.

4. Linear and non-linear correlation:

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

A	2	7	12	17
B	3	9	15	21

We can see that the relation of change between the variables is the same. If we plot these on the graph, we get straight line.

In a curvilinear or non-linear correlation, the amount of change in one variable does not bear a constant ratio of the amount of change in the other variables. The graph of non linear or curvilinear relationship will be a curve.

Methods of studying correlation:

There are 2 different methods for finding out the relationship between variables. They are

(1) Graphic Methods

a) Scatter diagram or Scattergram

b) Simple graph

(2) Mathematical Methods are

a) Karlpearson coefficient of correlation

b) Spearman's Rank correlation

c) Coefficient of concurrent deviation

d) Method of least squares

Scatter Diagram:

The scatter diagram is a chart obtained by plotting two variables to find out where there is any relation between them. In this diagram X variables are plotted on the horizontal axis and Y variables plotted on the vertical axis. Thus we can know the scatter or concentration of various points.

Various scatter diagrams are briefly shown below.

Advantage of scatter diagram:

1. Scatter diagram is a simple, attractive method to find out the nature of correlation.

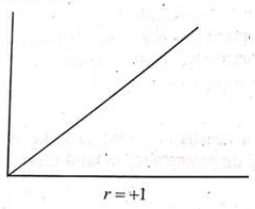
2. It is easy to understand

3. A rough idea is got at a glance whether it is positive or negative correlation.

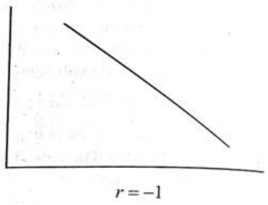
Simple Graph:

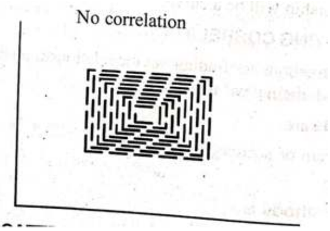
The values of the two variables are plotted on a graph paper. We get two curves, one for X variables another for Y variables. These two curves reveal the direction and closeness of the two variables and also reveal whether are not the variables are related. If both the curves move in the same direction, i.e., parallel to each other, either upward or downward, correlation is said to be positive. On the other hand, if they move in opposite directions, then the correlation is said to be negative.

Perfect positive correlation

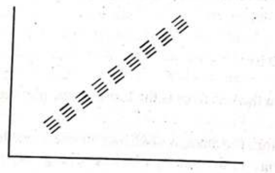


Perfect Negative correlation

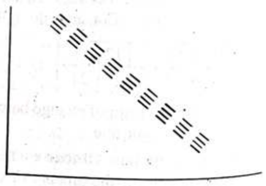




High degree of positive correlation



High degree of negitive correlation



Karl Pearson's correlation coefficient:

Correlation coefficient between two variables X and Y denoted by r_{XY} is a numerical measure of linear relationship between the two variables.

Limits of Karl Pearson's correlation coefficient:

The coefficient of correlation lies between $-1 \, and \, 1$ i.e., $-1 \le r_{XY} \le 1$.

If $r_{XY} = -1$, there is perfect negative correlation between 2 variables

$-1 < r_{XY} < 0$, variables are negatively correlated

$r_{XY} = 0$, there is no linear correlation

$0 < r_{XY} < 1$, variables are positively correlated

$r_{XY} = 1$, there is perfect positive correlation between 2 variables

Karl Pearson's correlation coefficient formulas:

We can find Karl Pearson correlation value by using following formulas.

1. $r_{XY} = \dfrac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$ Where, $x = X - \overline{X}$, $y = Y - \overline{Y}$ [x and y are large. Find deviations]

3. $r_{XY} = \dfrac{\text{cov}(XY)}{\sigma_X \ \sigma_Y}$

where $\text{cov}(XY) = \dfrac{1}{N}\sum (X - \overline{X})(Y - \overline{Y}) = \dfrac{1}{N}\sum XY - \overline{X}\overline{Y}$,

$$\sigma_X = \sqrt{\frac{1}{N}\sum (X - \overline{X})^2} = \sqrt{\frac{1}{N}\sum X^2 - (\overline{X})^2}$$

and $\sigma_Y = \sqrt{\dfrac{1}{N}\sum (Y - \overline{Y})^2} = \sqrt{\dfrac{1}{N}\sum Y^2 - (\overline{Y})^2}$

Example 1: Find if there is any significant correlation between the heights and weights given as

Height in inches	57	59	62	63	64	65	55	58	57
Weights in lbs	113	117	126	126	130	129	111	116	112

Solution: Coefficient of correlation $r = \dfrac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$

Computation of coefficient of correlation

Height in inches X	Deviation from Mean(60) $x = X - \overline{X}$	Square of deviations x^2	Weight in lbs Y	Deviation from Mean(120) $y = Y - \overline{Y}$	Square of deviations y^2	Product of deviation of x and y series xy
57	-3	9	113	-7	49	21
59	-1	1	117	-3	9	3
62	2	4	126	6	36	12
63	3	9	126	6	36	18
64	4	16	130	10	100	40
65	5	25	129	9	81	45
55	-5	25	111	-9	81	45
58	-2	4	116	-4	16	8
57	-3	9	112	-8	64	24
$\sum X = 540$	$\sum x = 0$	$\sum x^2 = 102$	$\sum Y = 1080$	$\sum y = 0$	$\sum y^2 = 472$	$\sum xy = 216$

$$N = 9,\ \overline{X} = \frac{540}{9} = 60 \quad and \quad \overline{Y} = \frac{1080}{9} = 120$$

$$r = \frac{216}{\sqrt{102 \times 471}} = 0.98$$

Example 2: Calculate Karl Pearson's correlation coefficient for the following paired data

x	28	41	40	38	35	33	40	32	36	33
y	23	34	33	34	30	26	28	31	36	38

What inference would you draw from the estimate.

Solution: Computation of coefficient of correlation

Height in inches X	Weight in lbs Y	X^2	Y^2	XY
28	23	784	529	644
41	34	1681	1156	1394
40	33	1600	1089	1320
38	34	1444	1156	1292
35	30	1225	900	1050
33	26	1089	676	858
40	28	1600	784	1120
32	31	1024	961	992
35	36	1225	1296	1260
33	38	1089	1444	1254
$\sum x = 355$	$\sum y = 313$	$\sum X^2 = 12761$	$\sum Y^2 = 9991$	$\sum XY = 11184$

Here N=10

$$r_{XY} = \frac{\text{cov}(XY)}{\sigma_X \ \sigma_Y} = 0.4133$$

Exercise:

1. Calculate Karl Pearson's coefficient of correlation from the following :

x	60	55	50	56	30	70	40	35	80	80	75
y	65	40	35	75	63	80	35	20	80	60	60

2. Calculate Karl Pearson's coefficient of correlation from the following :

x	9	8	7	6	5	4	3	2	1
y	15	16	14	13	11	12	10	8	9

3. Calculate coefficient of correlation between age of cars and annual maintenance costs and comment.

Age of cars	2	4	6	7	8	10	12
Annual maintenance cost	1600	1500	1800	1900	1700	2100	2000

4. Find coefficient of correlation from the following :

Height of father(inches)	65	66	67	68	69	71	73
Height of son(inches)	67	68	64	72	70	69	70

5. Given n=10, $\sigma_x = 4.5, \sigma_y = 3.6$ and the sum of product of deviation from the mean of X and Y is 64. Find the correlation coefficient.

6. On calculating correlation coefficient between two variables X and Y, a student got the following results: $n = 9, \sum X = 45, \sum Y = 64, \sum X^2 = 243, \sum Y^2 = 482, \sum XY = 340$. But, later when verified the problem, he found that data was entered wrong as

X	Y
8	10
9	12

instead of

X	Y
9	10
8	11

. What would be the correct correlation coefficient?

Answers:

1. 0.6084 2. 0.95 3. 0.836 4. 0.603 5.0.395 6. 0.833

Spearman Rank Correlation Coefficient:

A British Psychologist Charles Adward Spearman found out the method of finding the coefficient of correlation by ranks. This method based on rank and is useful in dealing with qualitative characteristics such as morality, character, intelligence and beauty. It cannot be measured quantitatively as in the case of Pearson's coefficient of correlation. It is based on the ranks given to the observations. Rank correlation is applicable only to the individual observations. The formula for Spearman's rank correlation is given by

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \quad i = 1,2,3,\dots,n$$

Where $\rho \rightarrow$ Rank coefficient of correlation

$d_i^2 \rightarrow$ Sum the of squares of the differences of two ranks

$n \rightarrow$ Number of paired observations

Properties of Rank Correlation Coefficient:

1. The limits of ρ lies between $-1\,and\,1$ i.e., $-1 \le \rho \le 1$
2. If $\rho = 1$, then there is perfect agreement in the order of the ranks and the direction of the rank is same.
3. If $\rho = -1$, then there is complete disagreement in the order of the ranks and they are in opposite directions.
4. If $\rho = 0$, then X and Y are independent.

Example 1: Following are the rank obtained by 10 students in two subjects, statistics and mathematics. To what extent the knowledge of the students in two subjects is related?

Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	2	4	1	5	3	9	7	10	6	8

Solution:

Rank in statistics (x)	Rank in Mathematics (y)	$d = (x-y)$	d_i^2
1	2	-1	1
2	4	-2	4
3	1	+2	4
4	5	-1	1
5	3	+2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	+3	9
10	8	+2	4
			$\sum d_i^2 = 40$

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6\times40}{10(10^2-1)} = 1 - \frac{240}{990} = 1 - 0.24 = 0.76$$

Example 2: A random sample of 5 college students is selected and their grades in 2 subjects are:

Subject 1	85	60	73	40	90
Subject 2	93	75	65	50	80

Solution:

Marks in subject 1	Ranks x	Marks in subject 2	Ranks y	Rank Differences $d_i = x - y$	d_i^2
85	2	93	1	1	1
60	4	75	3	1	1
73	3	65	4	-1	1
40	5	50	5	0	0
90	1	80	2	1	1
					4

Spearman's Rank correlation $= \rho = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6\times4}{5(5^2-1)} = 1 - \frac{1}{5} = 1 - 0.2 = 0.8$

Example 3: Ten competitors in a musical test were ranked by the three judges A, B and C as:

Ranks by A	1	6	5	10	3	2	4	9	7	8
Ranks by B	3	5	8	4	7	10	2	1	6	9
Ranks by C	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pair of judges has the nearest approach to common linking in music.

Solution: Here $N = 10$

Ranks by A (X)	Ranks by B (Y)	Ranks by C (Z)	$d_1 = X - Y$	$d_2 = X - Z$	$d_3 = Y - Z$	d_1^2	d_2^2	d_3^2
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	4
Total			0	0	0	200	60	214

$\rho_1(X,Y) = 1 - \frac{6\sum D_1^2}{n(n^2-1)} = 1 - \frac{6\times200}{10(99)} = \frac{-7}{33}$ $\quad \rho_2(X,Z) = 1 - \frac{6\sum D_2^2}{n(n^2-1)} = \frac{7}{11}$ $\quad \rho_3(Y,Z) = 1 - \frac{6\sum D_3^2}{n(n^2-1)} = \frac{-49}{165}$

Since $\rho_2(X,Z)$ is maximum, concluded that the pair of judges A and C has the nearest approach to common likings in music.

Exercise:

1. The ranks of 15 students in two subjects A and B are given below, the two numbers within the brackets denoting the ranks of the same student in A and B respectively are (1,10), (2,7), (3, 2), (4,6), (5,4), (6,8), (7,3), (8,1), (9,11), (10,15), (11,9), (12,5), (13,14), (14,12), (15,13). Use Spearman's formula to find rank correlation coefficient.

2. Ten Competitors in a music contest are ranked by three judges in the following order.

1st Judge:	1	6	5	10	3	2	4	9	7	8
2nd Judge:	2	4	3	1	8	9	7	6	10	5
3rd Judge:	6	4	9	8	1	2	3	10	5	7

Using Rank correlation method, determine which pair of judges has the nearest approach to common likings in music?

3. The rank correlation coefficient between the marks in 2 subjects is 0.8. The sum of squares of the differences between the ranks is 33. Find the number of students.

Answer: 1. 0.5142 2. -0.2121, -0.5757, 0.6363 3.10

EQUAL OR REPEATED RANKS

In the case of tied observations a common rank is assigned to each of the repeated observations. This rank is the average of the ranks which these observations would have been assigned, if they were different from each other by small quantity.

In this case a correction factor is applied to the formula ρ, where

$$C.F = \sum_{j=1}^{k} \frac{m_j(m_j^2-1)}{12}, \; m_j \rightarrow$$ the no:of repeated observations for the jth tie and

$$\rho = 1 - \frac{6\left[\sum d_i^2 + C.F\right]}{n(n^2-1)}, i = 1,2,3,\ldots,n$$

Example 1: From the following data calculate the rank correlation coefficient after making adjustment for tied ranks.

A	48	33	40	9	16	16	65	24	16	57
B	13	13	24	6	15	4	20	9	16	19

Solution: First we have to assign ranks to the variables.

A	Rank of A(x)	B	Rank of B (y)	d_i=Rank(x)-Rank (y)	d_i^2
48	3	13	5.5	-2.5	6.25
33	5	13	5.5	-0.5	0.25
40	4	24	1	-3	9.00
9	10	6	8.5	1.5	2.25
16	8	15	4	-4	16.00
16	8	4	10	-2	4.00
65	1	20	2	-1	1.00
24	6	9	7	-1	1.00
16	8	6	8.5	-0.5	0.25
57	2	19	3	-1	1.00
					$\sum d_i^2 = 41$

16 is repeated 3 times in X items hence $m_1 = 3$.

Since 13 and 6 are repeated twice in Y items; hence $m_2 = 2, m_3 = 2$.

$$\text{C.F}=\frac{3(3^2-1)}{12}+\frac{2(2^2-1)}{12}+\frac{2(2^2-1)}{12}=3$$

$$\rho = 1-\frac{6\left[\sum d_i^2 + C.F\right]}{n(n^2-1)} = 1-\frac{6(41+3)}{9(9^2-1)} = 0.733$$

Exercise:

1. Obtain the rank correlation coefficient between the variables X and Y from the following

A	50	55	65	50	55	60	50	65	70	75
B	110	110	115	125	140	115	130	120	115	160

2. Obtain the rank correlation coefficient between the variables X and Y from the following

X	42	64	42	42	60	45
Y	23	40	21	23	25	25

3. The following table gives the score obtained by 11 students in Science and Maths. Find the rank correlation coefficient

Social	40	46	54	60	70	80	82	85	85	90	95
Mathematics	45	45	50	43	40	75	55	72	65	42	70

4. Find the rank correlation coefficient between the heights of fathers and sons from the following data

Father	65	63	67	64	68	62	70	66	68	67	69	71
Son	68	66	68	65	69	66	68	65	71	67	68	70

Answer: 1. 0.1599 2. 0.85 3. 0.382 4. 0.722

Regression: Regression is a statistical method to estimate the unknown value of one variable from the known value of the related variable.

The lines described in the average relationship between two variables are known as lines of regression.

Comparison between Correlation and Regression:

Correlation	Regression
➢ It is a measure of degree of co variability between two variables ➢ In correlation, both the variables are random variables ➢ The coefficient of correlation is a relative measure	➢ Regression establishes the functional relationship between dependent and independent variables. ➢ In regression, one variable is dependent variable and other is independent ➢ Regression coefficient is an absolute measure

Linear Regression: If regression equation is a straight line then it is a linear regression, else non linear.
Lines of Regression: If the curve is a straight line of best fit for the given distribution of points is called the lines of regression.There are two such lines,
1. the best estimate values of Y for each specified value of X, is known as line of regression of Y on X.
2. the best estimate values of X for each specified value of Y, is known as line of regression of X on Y.

Fitting of Straight Line: Let $Y = a + bX$ be the straight line to be fitted to given set of points.
By Legendre Principle of least squares, a curve y=f(x) can be fitted to a given set of points (x, y) by minimizing the sum of squares of deviations of observed values from expected values as given by the curve of best fit.

$$E = \sum_{i=1}^{n} (a + bx_i - y_i)^2$$

To minimize $E, \frac{\partial E}{\partial a} = 0 \; \sum_{i=1}^{n} 2(a + bx_i - y_i) = 0 \Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$

and $\frac{\partial E}{\partial b} = 0 \; \sum_{i=1}^{n} 2(a + bx_i - y_i)(x_i) = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$

Regression equation of Y on X:
To determine the values of a and b for straight line $Y = a + bX$, the normal equations are

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Regression equation of X on Y:
To determine the values of a and b for straight line $X = a + bY$, the normal equations are

$$\sum X = na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

Example1: Determine equation of a straight line which best fit the data.
(or)
Find regression line on Y on X using least square method.

X	10	12	13	16	17	20	25
Y	10	22	24	27	29	33	37

Solution: Equation of straight line is $Y = a + bX$

Normal equations are $\sum Y = na + b\sum X$

$$\sum XY = a\sum X + b\sum X^2$$

X	Y	X^2	XY
10	10	100	100
12	22	144	264
13	24	169	312
16	27	256	432
17	29	289	493
20	33	400	660
25	37	625	925
$\sum X = 113$	$\sum Y = 182$	$\sum X^2 = 1938$	$\sum XY = 3186$

Substituting the values, we get

$$7a + 113b = 182 \ldots\ldots\ldots(1)$$

$$113a + 1983b = 3186 \ldots\ldots(2)$$

Solving eq (1) and (2), we get

$$a = 0.82 \text{ and } b = 1.56$$

Thus the equation of the straight line is

$$Y = a + bX$$

$$\therefore \quad Y = 0.82 + 1.56X$$

This is called the regression equation of Y on X.
Similarly we can find regression line of X on Y using $X = a + bY$

Exercise:

1. From a sample of 200pairs of observations the following quantities were calculated.

$$\sum X = 11.34, \quad \sum Y = 20.78, \quad \sum X^2 = 12.16, \quad \sum Y^2 = 84.96, \quad \sum XY = 22.13$$

From the above data, show how to compute the coefficients of the equation $Y = a + bX$.

2. Determine the equation of a straight line which best fits the data.

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

3. Fit a straight line to the following data.

X	1	2	3	4	5
Y	5	7	9	10	11

Answer: 1. Y=0.0007+1.819X 2. Y=11.9-0.65X (or) X=16.4-1.3Y 3. Y=3.9+1.5X

Deviations taken from Arithmetic mean of X and Y:

Regression line of X on Y is $X-\bar{X}=r_{XY}\dfrac{\sigma_X}{\sigma_Y}(Y-\bar{Y})$ (or) $X-\bar{X}=b_{XY}(Y-\bar{Y})$

Regression line of Y on X is $Y-\bar{Y}=r_{XY}\dfrac{\sigma_Y}{\sigma_X}(X-\bar{X})$ (or) $Y-\bar{Y}=b_{YX}(X-\bar{X})$

Two regression lines pass through the point $\left(\bar{X},\bar{Y}\right)$ where $b_{XY}=\dfrac{\sigma_X}{\sigma_Y}.r_{XY}$, $b_{YX}=\dfrac{\sigma_Y}{\sigma_X}.r_{XY}$

Properties of Regression Coefficient:

1. The correlation coefficient is the geometric mean of the regression coefficient.

$$\sqrt{b_{XY}.b_{YX}}=\sqrt{\frac{\sigma_X}{\sigma_Y}r\times\frac{\sigma_Y}{\sigma_X}r}=\sqrt{r^2}=\pm r$$

r, b_{XY}, b_{YX} all take same sign. i.e., if regression coefficients are positive then r is positive and if regression coefficients are negative then r is negative.

2. If one regression coefficient is greater than unity then the other must be less than unity.

Let $b_{XY}>1\Rightarrow\dfrac{1}{b_{XY}}<1$ then $r^2\le 1\Rightarrow b_{YX}.b_{YX}\le 1\Rightarrow b_{YX}\le\dfrac{1}{b_{XY}}<1\Rightarrow b_{YX}<1$

3. Arithmetic mean of the regression coefficient is $\geq$ correlation coefficient r (provided $r>0$)

To P.T $\dfrac{1}{2}(b_{XY}+b_{YX})\ge r\Rightarrow\dfrac{1}{2}\left(\dfrac{\sigma_X}{\sigma_Y}r+\dfrac{\sigma_Y}{\sigma_X}r\right)\ge r\Rightarrow\dfrac{\sigma_X}{\sigma_Y}+\dfrac{\sigma_Y}{\sigma_X}\ge 2$

$$\Rightarrow\sigma_X{}^2+\sigma_Y{}^2-2\sigma_X\sigma_Y\ge 0\Rightarrow(\sigma_X-\sigma_Y)^2\ge 0,\text{ which is true}$$

Example1: Find regression line of X on Y and Y on X. Also estimate weight when height is 67inches.

Height in inches	57	59	62	63	64	65	55	58	57
Weights in lbs	113	117	126	126	130	129	111	116	112

Solution:

Height in inches X	Deviation from Mean(60) $x=X-\bar{X}$	Square of deviations x^2	Weight in lbs Y	Deviation from Mean(120) $y=Y-\bar{Y}$	Square of deviations y^2	Product of deviation of x and y series xy
57	-3	9	113	-7	49	21
59	-1	1	117	-3	9	3
62	2	4	126	6	36	12
63	3	9	126	6	36	18
64	4	16	130	10	100	40
65	5	25	129	9	81	45
55	-5	25	111	-9	81	45
58	-2	4	116	-4	16	8
57	-3	9	112	-8	64	24
$\sum X=540$	$\sum x=0$	$\sum x^2=102$	$\sum Y=1080$	$\sum y=0$	$\sum y^2=472$	$\sum xy=216$

$$\overline{X}=\frac{540}{9}=60,\ \overline{Y}=\frac{1080}{9}=120 \qquad \overline{x}=\frac{0}{9}=0,\ \overline{y}=\frac{0}{9}=0$$

$$\sigma_X=\sqrt{\frac{1}{N}\sum x^2-(\overline{x})^2}=\sqrt{\frac{102}{9}-0}=\frac{\sqrt{102}}{3} \text{ and } \sigma_Y=\sqrt{\frac{1}{N}\sum y^2-(\overline{y})^2}=\sqrt{\frac{472}{9}-0}=\frac{2\sqrt{118}}{3}$$

Coefficient of correlation $r=\dfrac{COV(X,Y)}{\sigma_X\sigma_Y}=\dfrac{\frac{1}{N}\sum xy-\overline{x}\,\overline{y}}{\sigma_X\sigma_Y}=\dfrac{\frac{216}{9}-0}{\frac{\sqrt{102}}{3}\ \frac{2\sqrt{118}}{3}}=0.9844$

$$b_{XY}=\frac{\sigma_X}{\sigma_Y}.r_{XY}=\frac{\sqrt{102}/3}{2\sqrt{118}/3}\times 0.9844=0.4576 \qquad b_{YX}=\frac{\sigma_Y}{\sigma_X}.r_{XY}=\frac{2\sqrt{118}/3}{\sqrt{102}/3}\times 0.9844=2.1175$$

Regression line of X on Y is $X-\overline{X}=b_{XY}(Y-\overline{Y})\Rightarrow X-60=0.4576(Y-120)\Rightarrow X=0.4576Y-5.88$

Regression line of Y on X is $Y-\overline{Y}=b_{YX}(X-\overline{X})\Rightarrow Y-120=2.1176(X-60)\Rightarrow Y=2.1176X-7.056$

When $X=67$, $Y=2.1176(67)-7.056\Rightarrow Y=134.8232$

Example 2: Given the following data, calculate the expected value of Y when $X=12$

	x	y
Average	7.6	14.8
Standard deviation	3.6	2.5
$r=0.9$		

Solution: Regression line of Y on X is $Y-\overline{Y}=r_{XY}\dfrac{\sigma_Y}{\sigma_X}(X-\overline{X})$

$$Y-14.8=0.99\times\frac{2.5}{3.6}(X-7.6)\Rightarrow Y=0.688X+9.57$$

When X is 12; $Y=0.688(12)+9.57=17.826$

Example 3: Find the mean values of the variable X and Y and correlation coefficient from the following regression equations.

$2Y-X-50=0$ and $3Y-2X-10=0$

Solution:

$$2Y-X-50=0....(1) \qquad 3Y-2X-10=0...(2)$$

Solving (1) and (2) $Y=90$ and $X=130$. Since $(\overline{x},\overline{y})$ is a point on regression lines.

$\therefore \overline{x}=130$ and $\overline{y}=90$

Rewriting (1) and (2) $Y=\frac{1}{2}X+25$ *and* $X=\frac{3}{2}Y-5,\ \sigma=\frac{\sigma_X}{\sigma_Y}=\frac{3}{2}$ *and* $\sigma=\frac{\sigma_Y}{\sigma_X}=\frac{1}{2}$

$$\therefore\ r^2=\tfrac{3}{4}\Rightarrow r=0.866$$

Exercise:

1. Determine the regression equations for the following data:

X	2	4	6	8	10	12	14
Y	4	2	5	10	4	11	12

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

Price index of cotton	78	75	85	88	87	82	81	77	76	83	97	93
Price index of wool	84	82	82	85	89	90	88	92	83	89	98	99

3. The following data relate to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales in thousand rupees:

Salesmen Intelligence	A	B	C	D	E	F	G	H	I
Test Scores	50	60	50	60	80	50	80	40	70
Weekly Sales	30	60	40	50	60	30	70	50	60

(i) Obtain the regression equation of sales on intelligence test scores of the salesmen.
(ii) If the intelligence test score of a salesman is 65, what would be his expected weekly sales?

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

	Rainfall	Production
Average	30	500Kgs
Standard Deviation	5	100Kgs
Coefficient of correlation, r	0.8	

5. The following results were obtained in the analysis of data on yield of dry bark in ounces Y and age in years X of 200 cinchona plants. Find the two lines of regression and estimate the yield of dry bark of a plant of age 8 years.

	X	Y
Average	9.2	16.2
Standard Deviation	2.1	4.2
Coefficient of correlation, r	0.84	

6. Find the mean of the two series and the correlation coefficient.
For certain X and Y series which are correlated, the two lines of regression are:
(i) $7X-8Y+190=0,\ 16X-9Y-230=0$ (ii) $X=0.7Y+5.2, Y=0.3X+2.8$

7. If $b_{XY}=\dfrac{9}{20},\ b_{YX}=\dfrac{4}{5}$ and $\sigma_X=3$. Calculate (i) correlation coefficient (ii) S.D of Y

8. The regression of X on Y is 2Y-5X+160=0. If the mean value of Y is 40 and the variance of X is one fourth of the variance of Y, find the mean value of X and the correlation coefficient.

9. Using the data N=10, $\sum X=220, \sum Y=280, \sum\left(X-22\right)^2=140, \sum\left(Y-28\right)^2=108,$
$\sum(X-22)(Y-28)=81$, determine the correlation coefficient and regression equation of Y on X

10. Test whether 2X+3Y=4 and X-Y=5 represent valid regression lines.

Answers:

1. X=0.844Y+2.216, Y=0.732X+1.144   2. X=0.795Y+13.38, Y=0.59X+39.05
3. Y=0.75X+5, 53.75   4. 660Kgs   5. Y=1.68X+0.7, X=0.42Y+2.39, Y=14.02
6. (i) 54.615, 71.538, 0.7015   (ii)9.06, 5.519, 0.1   7. (i) 0.6 (ii) 4   8. 48, 0.8
9. 0.6587, Y=0.578X+15.27   10. No

Computer Oriented Statistical Methods (A8005)
#3.2 Fundamental Sampling Distributions

Population and Sample: Population is the aggregate or totality of statistical data forming a subjective of investigation. A finite subset of the population is Sample.

For example,
(i) the population of the heights of Indians
(ii) the population of Nationalized banks in India

Sampling: The process of partial enumeration is known as Sample survey. The results are then generalized and made applicable to the whole field of enquiry is known as Sampling.

Random Sampling:

The process of drawing a sample from a population in such a way that each member of the population has an equal chance of being included in the sample. The sample obtained by the process of random sampling is called a random sample.

For example,
(i) selecting randomly 20 words from a dictionary
(ii) choosing 10 patients from a hospital in order to test the efficacy of a certain newly-invented drug.

If each element of a population may be selected more than once then it is called **sampling with replacement** where as if the element cannot be selected more than once, it is called **sampling without replacement**.

Note:1. If N is the size of a population and n is the sample size, then
(i) The no:of samples with replacement = N^n
(ii) The no:of samples without replacement = ${}^{N}C_n$

2. If the size of the sample (n)<30, the sample is said to be small sample.

Parameters and Statistics: parameter is a statistical measure based on all the units (or observations) of a population. μ and σ are population mean and population S.D.

Statistic (or sample statistic) is statistical measure based on all units selected in a sample. $\overline{x}$ and s are sample mean and sample S.D.

Sample Mean and Sample Variance: If $x_1, x_2, x_3, \ldots., x_n$ represent a random sample of size n, then

Sample mean, $\overline{x} = \frac{1}{n}\sum_{i=i}^{n} x_i$ and Sample variance, $s^2 = \frac{1}{n-1}\sum_{i=i}^{n}\left(x_i - \overline{x}\right)^2$

Note: s^2 is essentially defined to be the average of the squares of the deviations of the observations from their mean divided by $n-1$ not n.

Sampling Distribution: Sampling distribution of a statistic helps to get information about the corresponding population parameter.

Central limit Theorem:

If $\overline{x}$ is the mean of the random sample of size n taken from a population having the mean μ and a finite variance σ^2, then $z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ as $n \to \infty$ is approaches that of standard normal distribution.

Standard Error (S.E) of a statistic: The standard error of a statistic (i.e., S.E. of sample mean, or sample S.D) is the standard deviation of the sampling distribution of the statistic. The S.E. of sample mean $\overline{x}$ is the S.D of the sampling distribution of sample mean.

Formula for S.E:

1. S.E. of sample mean $\overline{x} = \frac{\sigma}{\sqrt{n}}$

2. S.E. of sample proportion $p = \sqrt{\frac{PQ}{n}}, Q = 1 - P$

3. S.E. of the difference of two sample means $\overline{x_1}$ and $\overline{x_2}$, $\left(\overline{x_1} - \overline{x_2}\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

4. S.E. of $\left(p_1 - p_2\right) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$, where p_1 and p_2 are the proportions of two random samples of sizes n_1 and n_2 drawn from two populations with proportions P_1 and P_2 respectively

For a finite population of size N, when a sample is drawn without replacement

(i) S.E of sample mean = $\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

(ii) S.E of sample proportion = $\sqrt{\frac{PQ}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

Sampling Distribution of Mean:

The probability distribution of $\overline{X}$ is called the sampling distribution of means. The sampling distribution of a statistic depends on the size of the population, the size of the samples and the method of choosing the samples.

Infinite Population: Suppose the samples are drawn from an infinite population (or) sampling is done **with replacement**, then

the mean of the sampling distribution of means,

$$\mu_{\overline{X}} = E\left[\overline{X}\right] = \frac{E\left[X_1+X_2+X_3+\ldots+X_n\right]}{n} = \frac{\mu+\mu+\mu+\ldots\ldots+\mu}{n} = \frac{n\mu}{n} = \mu$$

and variance, $\sigma_{\overline{X}}^{\ 2} = Var\left[\overline{X}\right] = V\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n^2}V\left[X_1 + X_2 + X_3 + \ldots X_n\right]$

$$= \frac{1}{n^2}V\left[\sigma_1^{\ 2} + \sigma_2^{\ 2} + \sigma_3^{\ 2} + \ldots + \sigma_n^{\ 2}\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Hence $\mu_{\overline{X}} = \mu$, $\sigma_{\overline{X}}^{\ 2} = \frac{\sigma^2}{n}$ and S.D of mean, $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

Finite Population: Consider a finite population of size N with mean μ and standard deviation σ. Draw all possible samples of size *n* **without replacement**, from this population.

Then the mean of the sampling distribution of means (for *N*>*n*) is $\mu_{\overline{X}} = \mu$

The variance is $\sigma_{\overline{X}}^{\,2} = \frac{\sigma^2}{n}.\left(\frac{N-n}{N-1}\right)$ and S.D is $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.\sqrt{\left(\frac{N-n}{N-1}\right)}$

Here, the factor $\left(\frac{N-n}{N-1}\right)$, is called the finite population **correction factor**.

Example 1: What is the correction factor if *n*=5 and *N*=200.

Solution: Given *N*= the size of the finite population = 200, *n*= the size of the sample = 5

Correction factor = $\frac{N-n}{N-1} = \frac{200-5}{200-1} = \frac{195}{199} = 0.98$

Example 2: A population consists of five numbers 2, 3, 6, 8 and 11. Consider all possible samples of size two which can be drawn with replacement from this population. Find

(a) The mean of the population.
(b) The S.D of the population.
(c) The mean of the sampling distribution of means and
(d) The S.D of the sampling distribution of means

Solution: (a) Mean of the population $\mu = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6$

(b)Variance of the population is the given by,

$$\sigma^2 = \sum\frac{\left(x_i-\overline{x}\right)^2}{n} = \frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5} = 10.8$$

$$\sigma = \sqrt{10.8} = 3.29$$

(c)Mean of Sampling distribution of means **with** replacement:

The total no.of samples with replacement is $N^n = 5^2 = 25$ samples of size 2.

Here *N* = population size and *n*= sample size listing possible sample of size 2 from population 2, 3, 6, 8, 11 with replacement we get 25 samples

$$\left\{\begin{array}{lllll}
(2,2) & (2,3) & (2,6) & (2,8) & (2,11) \\
(3,2) & (3,3) & (3,6) & (3,8) & (3,11) \\
(6,2) & (6,3) & (6,6) & (6,8) & (6,11) \\
(8,2) & (8,3) & (8,6) & (8,8) & (8,11) \\
(11,2) & (11,3) & (11,6) & (11,8) & (11,11)
\end{array}\right\}$$

The arithmetic means for each of these 25 samples. The sample means are

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

The mean of sampling distribution of means is the mean of these 25 means.

$$\mu_{\overline{X}}=\frac{\text{Sum of all sample means}}{25}=\frac{150}{25}=6=\mu$$

(d)S.D of Sampling distribution of means **with** replacement

$$\sigma_{\overline{x}}{}^{2}=\frac{(2-6)^2+(2.5-6)^2+(4-6)^2+......+(11-6)^2}{25}=5.40$$

and thus $\sigma_{\overline{x}}=\sqrt{5.40}=2.32$

Note: Clearly $\sigma_{\overline{X}}{}^{2}=\frac{\sigma}{\sqrt{n}}=\frac{3.29}{\sqrt{2}}=2.32$

Example 3: A population consists of five numbers 2, 3, 6, 8 and 11. Consider all possible samples of size two which can be drawn without replacement from this population. Find

(a) The mean of the population
(b) The S.D of the population.
(c) The mean of the sampling distribution of means and
(d) The S.D of the sampling distribution of means

Solution:

(a) Mean of the population $\mu=\frac{2+3+6+8+11}{5}=\frac{30}{5}=6$

(b) Variance of the population is the given by,

$$\sigma^2=\sum\frac{(x_i-\overline{x})^2}{n}=\frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5}=10.8$$

$$\sigma=\sqrt{10.8}=3.29$$

(c) Mean of Sampling distribution of means **without** replacement:

The total no.of samples with replacement is ${}^{N}C_{n}={}^{5}C_{2}=10$ samples of size 2.

The 10 samples are

$$\left\{\begin{array}{llll}(2,3) & (2,6) & (2,8) & (2,11)\\ (3,6) & (3,8) & (3,1) & \\ (6,8) & (6,11) & & \\ (8,11) & & & \end{array}\right\}$$

The corresponding sample means are

$$\left\{\begin{array}{llll}2.5 & 4 & 5 & 6.5\\ 4.5 & 5.5 & 7 & \\ 7 & 8.5 & & \\ 9.5 & & & \end{array}\right\}$$

The mean of sampling distribution of means is

$$\mu_{\overline{X}}=\frac{2.5+4+5+6.5+4.5+5.5+7+7+8.5+9.5}{10}=6=\mu$$

(d) S.D of Sampling distribution of means **without** replacement

$$\sigma_{\overline{x}}{}^{2}=\frac{(2.5-6)^2+(4-6)^2+(5-6)^2+......+(9.5-6)^2}{10}=4.05$$

and thus $\sigma_{\overline{x}}=\sqrt{4.05}=2.01$

Note: Clearly $\sigma_{\overline{X}}{}^{2}=\frac{\sigma^2}{n}.\left(\frac{N-n}{N-1}\right)=\frac{10.8}{2}.\left(\frac{5-2}{5-1}\right)=4.05$

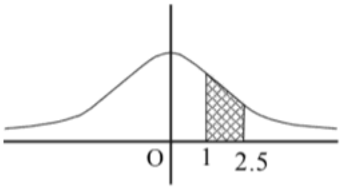
Example 4: A random sample of size 64 is taken from an infinite population having the mean 45 and S.D 8. What is the probability that $\bar{x}$ will be between 46 and 47.50.

Solution: Given $n = 64, \mu = 45, \sigma = 8$

The standard normal variate corresponding to $\bar{x}$ is $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

when $\bar{x} = 46$, $z_1 = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{46-45}{8/\sqrt{64}} = 1$

when $\bar{x} = 47.5$, $z_1 = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{47.5-45}{8/\sqrt{64}} = 2.5$



Required probability, $P(46 < \bar{x} < 47.5) = P(1 < z < 2.5)$

$= P(0 < z < 2.5) - P(0 < z < 1) = 0.4938 - 0.3413 = 0.1525$

Exercise:

1. Find the value of the finite population correction factor for n=10 and N=100. [Ans: 0.991]
2. How many different samples of size two can be chosen, from a finite population of size 25. [Ans:300]
3. The variance of a population is 2. The size of the sample collected from the population is 169. What is the standard error of the mean. [Ans: 0.108]
4. A population consists of six numbers 5, 10, 14, 18, 13 and 24. Consider all possible samples of size two which can be drawn without replacement from this population. Find
   (a) The mean of the population.
   (b) The S.D of the population.
   (c) The mean of the sampling distribution of means and
   (d) The S.D of the sampling distribution of means [Ans: 14, 5.972, 14, 3.78]

5. A population consists of six numbers 4, 8, 12, 16, 20 and 24. Consider all possible samples of size two which can be drawn without replacement from this population. Find
   (a) The mean of the population.
   (b) The S.D of the population.
   (c) The mean of the sampling distribution of means and
   (d) The S.D of the sampling distribution of means [Ans: 14, 6.83, 14, 4.32]

6. A population consists of five numbers 3, 6, 9, 15 and 27. Consider all possible samples of size two which can be drawn with replacement from this population. Find
   (a) The mean of the population.
   (b) The S.D of the population.
   (c) The mean of the sampling distribution of means and
   (d) The S.D of the sampling distribution of means [Ans: 12, 8.485, 12, 6.03]

7. If the population is 3, 6, 9, 15 and 27
   (a) list all possible samples of size 3 that can be taken without replacement from the finite population
   (b) Calculate the mean of each of the sampling distribution of means
   (c) Find the S.D of sampling distribution of means. [Ans: (b) 12 (c)3.651]

8. The mean height of students in a college is 155 cms and S.D is 15. What is the probability that mean height of 36 students is less than 157cms. [Ans: 0.7881]

9. A random sample of size 100 is taken from an infinite population having the mean 76 and variance 256. What is the probability that $\overline{x}$ will be between 75 and78. [Ans: 0.628]

10. If a 1-gallon can paint covers on an average 513 sq.feet with a S.D of 31.5sq.feet, what is the probability that the mean area covered by a sample of 40 of these 1-gallon cans will be anywhere from 510 to 520 sq.feet? [Ans:0.645]

11. If the mean of breaking strength of copper wire is 575 lbs, with a S.D 8.3lbs. How large a sample must be used in order that there will be one chance in 100 that the mean breaking strength of the sample is less than 572lbs? [Ans: n = 42]

12. The guaranteed average life of a certain type of electric bulbs is 1500hrs with a S.D of 120hrs. It is decided to sample the output so as to ensure that 95% of bulbs do not fall short of the guaranteed average by more than 2%. What will be the minimum sample size? [Ans: n = 44]

13. A normal population has a mean of 0.1 and S.D of 2.1. Find the probability that mean of a sample size 900 will be negative. [Ans: 0.0764]

14. Determine the expected number of random samples having their means (a) between 22.39 and 22.41 (b) greater than 22.42 (c) less than 22.37 (d) less than 22.38 and more than 22.41 for the following data.
N = size of the population = 1500, n = size of the sample = 36, Number of samples (N) =300,
σ = population S.D = 0.48, μ = Population mean = 22.4 [Ans: 28, 120, 1, 255]

Sampling Distribution of the mean (σ unknown):

For small sample of size (n<30), when σ is unknow, it can be substituted by *s*, provided the sample is drawn from a normal population. The probability distributions used are

1. Student's t-distribution
2. Chi-square distribution
3. Snedecor's F-distribution

Degrees of Freedom:

The number of independent variates which make up the statistic is known as the degrees of freedom (d.f) and is denoted by ν (Nu). It is a number which indicates how many of the values may be independently (or freely) chosen.

In general, the no.of degrees of freedom is equal to the total number of observations less the no.of independent constraints imposed on the observations. In a set of data of *n* observations, if *k* is the no.of independent constraints then $\nu = n - k$.

Student's t-Distribution or t-Distribution:

It is used for testing the hypothesis when the sample size is small and population S,D σ is unknown.

The test statistic t is defined by $t = \frac{\bar{x}-\mu}{S/\sqrt{n}} \sim n-1$ d.o.f,

where $\bar{x}$ = sample mean and $S^2 = \frac{1}{n-1}\sum_{i=i}^{n}\left(x_i - \bar{x}\right)^2$

Chi-square $\left(\chi^2\right)$ Distribution:

It is used as a measure of goodness of fit and to test the independence of attributes.

Let S^2 be the variance of a random sample of size *n*, taken from a normal population having the variance σ^2. Then

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=i}^{n}\left(x_i-\bar{x}\right)^2}{\sigma^2} \sim n-1 \text{ d.o.f.} \quad \left[\because S^2 = \frac{1}{n-1}\sum_{i=i}^{n}\left(x_i-\bar{x}\right)^2\right]$$

F- Distribution (Sampling distribution of the ratio of two sample variances):

Let $S_1^{\,2}$ and $S_2^{\,2}$ be two variances of two random samples of sizes n_1 and n_2 respectively drawn from normal population, under the assumption that two normal populations have same variances $\left(\sigma_1^{\,2} = \sigma_2^{\,2}\right)$,

We have $F = \frac{S_1^2}{S_2^2} \sim (n_1 - 1, n_2 - 1)$ d.o.f. (provided $S_1^{\,2} > S_2^{\,2}$)

Example 1: Find (a) $t_{0.05}$ when $\nu = 16$ (b) $t_{-0.01}$ when $\nu = 10$ (c) $t_{0.995}$ when $\nu = 7$

Solution: (a) when $\nu = 16,\ t_{0.05} = 1.746$

(b) when $\nu = 10, t_{-0.01} = -2.764$

(c) when $\nu = 7, t_{0.995} = t_{1-0.005} = -t_{0.005} = -3.499$

Example 2: For an *F*-Distribution, Find

(a) $F_{0.05}$ when $\nu_1 = 7$ and $\nu_2 = 15$ (b) $F_{0.95}$ when $\nu_1 = 19$ and $\nu_2 = 24$

Solution: (a) when $\nu_1 = 7$ and $\nu_2 = 15, F_{0.05} = 2.71$

(b) $F_{0.95}(19,24) = \frac{1}{F_{0.05}(19,24)} = \frac{1}{2.11} = 0.4739$

Example 3: A random sample of size 25 from a normal population has the mean $\bar{x}$ =47.5 and S = 8.4. Does this information tend to support or refuse the claim that the mean population is μ =42.5? (For 24 d.o.f, with α =0.05, *t*=2.797 from t-distribution table)

Solution: Given *n* = 25, $\bar{x}$ =47.5, S = 8.4 and μ =42.5

We have *t*-distribution, $t = \frac{\bar{x}-\mu}{S/\sqrt{n}} = \frac{47.5-42.5}{8.4/\sqrt{25}} = 2.98 \sim (25-1 = 24\,d.o.f\,)$

Since 2.98 > 2.797(table value), refuse the claim that the mean population is μ =42.5

Example 4: A manufacture claims that any of his list of items cannot have variance more than 1 cm^2. A sample of 25 items has a variance of 1.2 cm^2. Test whether the claim of the manufacturer is correct. (For 24 d.o.f, with α =0.05, χ^2 =36.41 from distribution table)

Solution: Given n = 25, S^2 = 1.2 and σ^2 =1

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{24(1.2)^2}{1^2} = 28.8(\sim 25-1=24\, d.o.f)$$

Clearly 28.8 < 36.41 (table value). The claim is correct.

Example 5: If two independent random samples of sizes 13 and 17 are taken in a normal population. What is the probability that the variance of the first sample will be at least four times as large as that of the second sample? (For (12, 16) d.o.f, with α =0.05, F = 4.00 from distribution table)

Solution: Given $n_1 = 13$ and $n_2 = 7$, $S_1^2 = 4S_2^2$

$$F = \frac{S_1^2}{S_2^2} = \frac{4S_2^2}{S_2^2} = 4 \sim (13-1, 7-1)$$

Since F = 4 (same as table value). The required probability is 0.05.

Exercise:

1. For an F-Distribution, Find

(a) $F_{0.05}$ when $\nu_1 = 15$ and $\nu_2 = 7$ (b) $F_{0.95}$ when $\nu_1 = 12$ and $\nu_2 = 15$

(c) $F_{0.95}$ when $\nu_1 = 10$ and $\nu_2 = 20$ [Ans: 3.51, 0.38, 0.36]

2. A process for making certain ball bearings is under control if the diameters of the bearings have a mean of 0.5000 cm. If a random sample of 10 of these bearings has a mean diameter of 0.5060 cm and S.D of 0.0040 cm, is the process under control? [Ans: t=4.7334, not under control] (For 9 d.o.f, with α =0.05, t=3.250 from t-distribution table)

3. A sample of 20 parts used in a semiconductor has a variance of 0.84, thousand of a inch. Test whether the process is under control if population variance is 0.60, thousands of an inch (For 19 d.o.f, with α =0.05, χ^2 =36.191 from distribution table) [Ans: 37.24, not under control]

5. If two independent random samples of sizes $n_1 = n_2 = 8$ are taken in a normal population having the same variances. What is the probability that the variance of the first sample will be at least seven times as large as that of the second sample? (For (7, 7) d.o.f, with α =0.05, F = 3.787 from distribution table)

COMPUTER ORIENTED STATISTICAL METHODS (A8005)
#4. Estimations and Testing of Hypothesis for Large Samples

Point Estimation, Maximum Error Estimate

Estimate: An estimate is a statement made to find an unknown population parameter.

Estimator: The procedure or rule to determine an unknown population parameter is called an estimator.

Types of Estimation:

Basically, there are two kinds of estimates to determine the statistic of the population parameters namely, (a) Point Estimation and (b) Interval Estimation.

Point Estimation and Interval Estimation:

If an estimate of the population parameter is given by a single value, then the estimate is called Point Estimation of the parameter.

If an estimate of a population parameter is given by two different values between which the parameter may be considered to lie, then the estimate is called an interval estimation of the parameter.

Example: If the height of a student is measured as 162 cm, then the measurement gives point estimation. But if the height is given as (163 ± 3.5) cm, then the height lies between 159.5 cm and 166.5cms and the measurement gives interval estimation.

Unbiased and biased estimates:

A statistic is said to be unbiased estimator of the corresponding parameter if the mean of the sampling distribution of the statistic is equal to the corresponding population parameter. Otherwise the statistic is called a biased estimator of the corresponding parameter .The value of statistics in the above two cases are called unbiased and biased estimates respectively.

If t be a statistic and θ be the corresponding parameter and E(t) = θ, then t is an unbiased estimator of θ and the bias is E(t) - θ.

Result: Sample mean $\overline{x}$ is an unbiased estimator of population mean μ.
Proof: Let $x_1, x_2 \ldots\ldots, x_n$ be a random sample drawn from a given population with mean μ and variance σ², then

$$E(\overline{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}E(x_1 + x_2 + x_3 + \ldots\ldots + x_n) \quad = \frac{1}{n}[E(x_1) + E(x_2) + E(x_3) + \ldots\ldots + E(x_n)]$$

$$= \frac{1}{n}[\mu + \mu + \mu + \ldots\ldots\ldots + \mu] \qquad [\because E(x_i) = \mu]$$

Hence the sample mean $\overline{x}$ is an unbiased estimator of the population mean μ.

Confidence Interval and Maximum error estimate:

Since the sample mean $\overline{x}$ estimate very rarely equals to the population mean μ

The error is $\left|\overline{X}-\mu\right|=E$

For large n, $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is the normal variate and $1-\alpha$ is degree of confidence, then

$$P(-z_{\alpha/2}<z<z_{\alpha/2})=(1-\alpha)100\% \Rightarrow P\left(-z_{\alpha/2}<\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}<z_{\alpha/2}\right)=1-\alpha$$

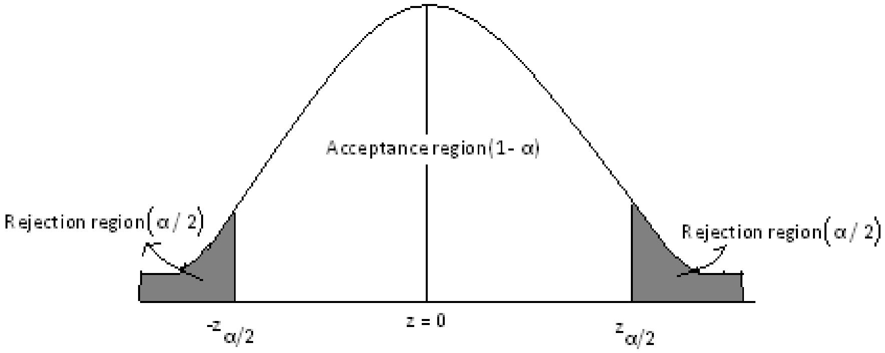
$\Rightarrow\left(\overline{x}-z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}<\mu<\overline{x}+z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right)=1-\alpha$ is confidence intertval

and maximum error is $E=\dfrac{z_{\alpha/2}.\sigma}{\sqrt{n}}$

Sample size: $n=\left(\dfrac{z_{\alpha/2}.\sigma}{E}\right)^2$

Confidence level:

Confidence interval is also known as acceptance region and critical region is known as rejection region



$P\left(-z_{\alpha/2}\le z\le z_{\alpha/2}\right)=(1-\alpha)100\%$

Note: If $\alpha=5\%$ (rjection region) then 1-$\alpha=95\%$ (acceptance region)

$1-\alpha=0.95$ then $P(-z_{\alpha/2}<z<0)=0.475$ and $P(0<z<z_{\alpha/2})=0.475$

from normal tables the area corresponding to 0.475 is $z_{\alpha/2}=1.96$

Similarly, at 90%, $z_{\alpha/2}=1.645$ and at 99%, $z_{\alpha/2}=2.58$

Confidence level, α	90	95	98	99
Critical value, $z_{\alpha/2}$	1.64	1.96	2.33	2.58

Example 1: In a study of an automobile insurance a random sample of 80 body repair costs had a mean of Rs.472.36 and the S.D of Rs.62.35.If $\overline{x}$ is used as appoint estimate to the true average repair costs, with what confidence we can assert that the maximum error doesn't exceed Rs.10.

Solution: Size of a random sample, n=80

The mean of random sample, $\bar{x}$ =Rs.472.36

Standard deviation, σ =Rs.62.35

Maximum error of estimate, $E_{\max}$ =Rs.10

We have $E_{\max} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow z_{\frac{\alpha}{2}} = \frac{E_{\max} \cdot \sqrt{n}}{\sigma} = \frac{10\sqrt{80}}{62.35} = \frac{89.4427}{62.35} = 1.4345$

$z_{\alpha/2} = 1.43$. Then area when $z = 1.43$ from tables is 0.4236

$\frac{\alpha}{2} = 0.4236 \Rightarrow \alpha = 0.8472$

Confidence = (1-α) 100%=84.72%

Hence we are 84.72% confidence that the maximum error is Rs.10.

Example 2: What is the maximum error one can expect to make with probability 0.95 when using the mean of a random sample of size n=64 to estimate the mean of population with variance 2.56.

Solution: Given Confidence limit =95%

i.e (1-α)100=95 then $z_{\alpha/2} = 1.96$

Maximum error estimate is given by $E = \frac{z_{\alpha/2} . \sigma}{\sqrt{n}} = \frac{1.96(1.6)}{\sqrt{64}} = \frac{49}{125} = 0.392$

Example 3: It is desired to estimate the mean number of hours of continuous use until a certain computer will first require repairs. If it can be assumed that $\sigma = 48$ hours, how large a sample be needed so that one will be able to assert with 90% confidence that the sample mean is off by at most 10 hours.

Solution: Given Maximum error=10 hours, $\sigma = 48$ Hours and $z_{\alpha/2} = 1.645$ (for 90%)

$$n = \left(\frac{z_{\frac{\alpha}{2}} . \sigma}{E} \right)^2 = \left(\frac{1.645 \times 48}{10} \right)^2 = 62.3 = 62$$

Hence sample size =62

EXERCISE

1. To estimate the average amount of time visitors take to move from one building to another in an office complex, the mean of a random sample of size n is used. Given $\sigma = 1.40$ minutes, determine how large should be the sample size if it is ascertained with 99% confidence that the error E is at most 0.25. [Ans: 209]

[**Hint:** Given $\sigma = 48$, E=0.25, $\alpha = 99\%$ then $z_{\alpha/2} = 2.58$, find n]

2. Using the mean of random sample of size 150 to estimate the mean mechanical aptitude of mechanics of a large workshop and assuming$\sigma = 6.2$, what can we assert with 0.99 probabilities about the maximum size of the error. [Ans: 1.30]

[**Hint:** Given $\sigma = 6.2$, n=150, $\alpha = 99\%$ then $z_{\alpha/2} = 2.58$, find E]

3. It is desired to estimate the mean time of continuous use until an answering machine will first require service. It it can be assumed that s.d is 60 days, how large a sample is needed so that one will be able to assert with 90% confidence that the sample mean is off by at most 10 days.
[**Hint:** Given $\sigma = 60$, E=10, $\alpha = 90\%$ then $z_{\alpha/2} = 1.64$, find n] [Ans: 97]

4. The research worker wants to determine the average time it takes a machine to rotate the tyres of a car and he wants to be able to assert with 95% Confidence that the mean of his sample is off by at most 0.5 minutes. If he can presume from past experience that $\sigma = 1.6$ minutes, how large a sample will he have to take? [Ans: 46]
[**Hint:** Given $\sigma = 1.6$, E=0.5, $\alpha = 95\%$ then $z_{\alpha/2} = 1.96$, find n]

5. Find the degree of confidence to assert that the average salary of school teachers is between Rs.272 and Rs.302 if a random sample of 100 such teachers revealed a mean salary of Rs.287 with S.D of Rs.48. [Ans: 99.82%]

Solution: X: salary of teachers

$$\mu = 287, \quad \sigma = 48, n = 100$$

$$X_1 = 272, \; z_1 = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{272 - 287}{48 / \sqrt{100}} = -3.125$$

$$X_2 = 302, \; z_1 = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{302 - 287}{48 / \sqrt{100}} = 3.125$$

$$\therefore P(272 \le X \le 302) = P(-3.125 \le z \le 3.125) = 2P(3.125) = 2(.4991) = 0.9982$$

[From normal tables for z=3.12, area=0.4991]

The degree of confidence is $\alpha = 99.82\%$

CONFIDENCE INTERVAL ESTIMATES OF PARAMETERS

The formulas for confidence limits of some well-known statistic for large random samples are:

(a) **Confidence limits for population mean** μ

Standard Error of $\bar{x}$ is given by $\frac{\sigma}{\sqrt{n}}$

Confidence limits are $\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$

(b) **Confidence limits for population proportion P**

Standard Error of p is given by $\sqrt{\frac{pq}{n}}$

Confidence limits are $\left(p - z_{\alpha/2} \sqrt{\frac{pq}{n}}, p + z_{\alpha/2} \sqrt{\frac{pq}{n}} \right)$

(c) **Confidence limits for difference** $\mu_1 - \mu_2$ **of two population means** μ_1 **and** μ_2

Standard Error of $\left(\overline{x}_1 - \overline{x}_2\right)$ is given by $\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

Confidence limits are $\left(\left(\overline{x}_1 - \overline{x}_2\right) - z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}, \left(\overline{x}_1 - \overline{x}_2\right) + z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}\right)$

(d**) Confidence limits for the difference** $P_1 - P_2$ **of two population proportions**

Standard Error of $\left(p_1 - p_2\right)$ is given by $\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}$

Confidence limits are $\left(\left(p_1 - p_2\right) - z_{\alpha/2}\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}, \left(p_1 - p_2\right) + z_{\alpha/2}\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}\right)$

Example1: The mean and the standard deviation of a population are 11,795 and 14,054 respectively. If n=50, find 95% confidence interval for the mean.

Solution: Here mean of population, $\mu = 11795$

S.D of population, $\sigma = 14054$

$\overline{x} = 11795$, n=sample size=50, $z_{\frac{\alpha}{2}}$ for 95% confidence =1.96

95% Confidence interval= $\left(\overline{x} - z_{\frac{\alpha}{2}} \cdot \dfrac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \dfrac{\sigma}{\sqrt{n}}\right)$

= (11795-3899, 11795+3899) = (7896, 15694)

Example 2: In a random sample of 400 industrial accidents, it was found that 231 were due to unsafe working conditions. Find (a) The maximum error (b) Construct 95% confidence interval

Solution: Here proportion of population, $p = \dfrac{231}{400}$ $\quad q = 1 - p = 1 - \dfrac{231}{400} = \dfrac{169}{400}$

n =sample size=400, $z_{\frac{\alpha}{2}}$ for 95% confidence =1.96

95% Confidence interval for population proportion is given by

$$\left(p - z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}}, p + z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}}\right)$$

$= (0.5775 - 0.0484, 0.5775 + 0.0484) = (0.5291, 0.6259)$

Example 3: In a certain factory there are two independent processes manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 ozs with standard deviation of 12 ozs. While the corresponding figures in a sample of 400 items from the other process are 124 and 14. Obtain the standard error of difference between the two sample means. Find the 99 % confidence limits for the difference in the average weights of items produced by the two processes respectively?

Solution: $n_1 = 250$ $\quad \bar{x}_1 = 120$ $\quad \sigma_1 = 12$ $\quad$ and $n_2 = 400$ $\quad \bar{x}_2 = 124$ $\quad \sigma_2 = 14$

$z_{\frac{\alpha}{2}}$ for 99% confidence = 2.58

99% Confidence limits are $\left(\overline{x_1} - \overline{x_2}\right) \pm 2.58\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

$$= 4 \pm 2.58(1.034) = (1.33, 6.67)$$

Example 4: A random sample of 300 shoppers at a super market includes 204 who regularly use cents off coupon. In another sample of 500 shoppers at a super market includes 75 who regularly use cents off coupon. Construct a 98 % confidence interval that any one shopper at the supermarket selected at random will regularly use cents off coupons.

Solution: $n_1 = 300 \qquad x_1 = 204 \qquad p_1 = \frac{x_1}{n_1} = \frac{204}{300} = 0.68$

$n_2 = 500 \qquad x_2 = 75 \qquad p_2 = \frac{x_2}{n_2} = \frac{75}{500} = 0.15$

98 % confidence limits of the difference of proportions is

$$(p_1 - p_2) \pm 2.33\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$= (0.68 - 0.15) \pm 2.33\sqrt{\frac{0.68 \times 0.32}{300} + \frac{0.15 \times 0.85}{500}}$$

$$= (0.53) \pm 2.33(0.031) = 0.53 \pm 0.072 = (0.458, 0.602)$$

Exercise

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

2. The mean and S.D of population are 11,795 and 14,054 respectively. If n=50, find 95% confidence interval for mean [Ans: (7899 , 15690)]

3. In a random sample of 160 workers exposed to a certain amount of radiation, 24 experienced some ill effects. Construct a 99% confidence interval for the corresponding true percentage.

4.100 articles from a factory are examined and 10 are found to be defective. 500 similar articles from a second factory are found to be 15 defective. Obtain the confidence level for the differences of two proportions at 5% level.

5. In a random sample of 100 packages shipped by air freight 13 had some damage. Construct 95% confidence interval for the true proportion of damage package.

6. A random sample of 500 pineapples was taken from a Large consignment and 65 were found to be bad. Obtain 98% confidence limits for the percentage of bad apples in the consignment.

7. Among 100 fish caught in a lake, 18 were inedible due to pollution. With what confidence can we assert that the error of this estimate is at most 0.065?

Introduction to Hypothesis:

We can find either a single number for the parameter or an interval of values. However, there are many problems, in which, rather than estimating the value of a parameter we need to decide whether to accept or reject a statement about the parameter. This statement is called a hypothesis and the decision-making procedure about the hypothesis is called hypothesis testing. This is one of the most useful aspects of statistical inference, since many types of decision-making, tests or experiments in the engineering world can be formulated as hypothesis –testing problems.

Hypothesis testing main aim is to provide rules that lead to decision resulting in acceptance (or) rejection of statements about the population parameter.

When parametric values are unknown, we estimate them through sample values. But the problem arises when the sample provides a value, which is neither exactly equal to the parametric value, nor too far.

So a procedure is developed which enables one to decide whether to accept a value or not on the basis of sample values. Such a procedure is known as **Testing of Hypothesis**.

Test of Statistical Hypothesis:

A statistical hypothesis is an assumption (or) procedure which makes one to decide about the acceptance (or) rejection of the hypothesis, denoted by H

Hypothesis is of two types:

(1) Null Hypothesis (2) Alternative Hypothesis

I Null Hypothesis:

Null Hypothesis denoted by H_o is the statistical hypothesis which is to be actually tested for acceptance (or) rejection.

In case of single statistic H_o : The sample statistic does not differ significantly from the parameter.

In case of two statistics H_o : The two sample statistics does not differ significantly.

Example 1:

In a factory, bulbs are manufactured under some process having an average life of μ hours. And if it is proposed to test a new procedure for manufacturing bulbs having an average life of μ_o hours.

There are two population of bulbs, those manufactured by standard process and those manufactured by new process.

Set up the Null hypothesis as,

H_o :There is no difference between the two process ie., $H_o : \mu = \mu_o$

II ALTERNATIVE HYPOTHESIS:

Any hypothesis which is complementary to the null hypothesis is called an Alternative Hypothesis, denoted by H_1

Example 2:

For the Example1, we can denote any one of the following alternative hypothesis as

(i) $H_1 : \mu \neq \mu_o$, there is a difference between the two processes
This case is known as two tailed alternative

(ii) $H_1 : \mu > \mu_o$, New process is inferior to standard process.
This case is known as right tailed alternative

(iii) $H_1 : \mu < \mu_o$, New process is superior to standard process.
This case is known as left tailed alternative

Errors in testing the Hypothesis:

The decision to accept (or) reject the null hypothesis Ho is made on the basis of the information supplied by the observed sample.

The four possible situations that arise in any test procedure

		Decision from sample	
		Reject H_o	Accept H_o
True Statement	Ho True	Wrong Decision **[Type I error]**	Correct Decision
	H_1 True (Ho False)	Correct Decision	Wrong Decision [**Type II error**]

In testing of hypothesis, there are two types of errors:

1) Type I error (or) First kind of error:

Reject H_o when H_o is true

This error is also known as Rejection error.

P[Type I error] = P[Reject H_o / H_o is true] = α

Example :

If a medicine is administered to few patients of a particular disease to cure them and medicine is curing the disease. But it is claimed that the medicine has no effect and hence the medicine is discontinued. This is ***Type I error***

2) Type II error (or) Second kind of error:

Accept H_o when H_o is false

This error is also known as Acceptance error.

P[Type II error] = P[Accept H_o / H_o is false] = β

Example 4:

For the example 3, The medicine is not curing the disease ie., it has adverse effect. But it is claimed to have good effect and the treatment is continued. This is ***Type II error***.

Critical Region:

A region corresponding to a statistic, in the sample space which leads to the rejection of H_o is called ***Critical region (or) rejection region***. The region which leads to acceptance of H_o is called ***Acceptance region***.

Critical values or significant values:

The value of the test statistic, which separates the critical region and the acceptance region is called the ***critical value (or) significant value,*** z_α . This value is dependent on

(1) The level of significance used

(2) The alternative hypothesis (whether it is one tailed (or) two tailed)

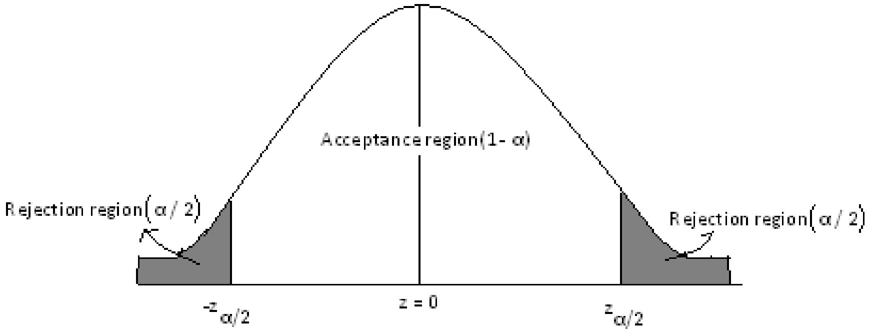
Level of significance :

The P[type I error] = α is known as level of significance. The level of significance usually employed in testing of hypothesis is 5% and 1% α which is always fixed in advance before collecting the sample information.

Two-tailed test at level of significance α:

The critical value z_α of the test statistic at level of significance α for a two tailed test is given by

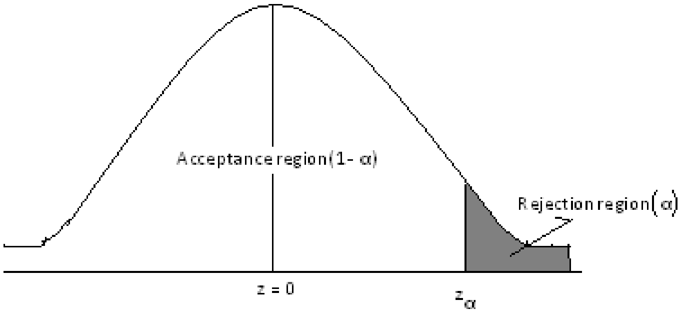
$$P\left(-z_{\alpha/2} < z < z_{\alpha/2}\right) = 1 - \alpha \text{(acceptance region)} \quad \text{------------(1)}$$



One-tailed test at level of significance α:

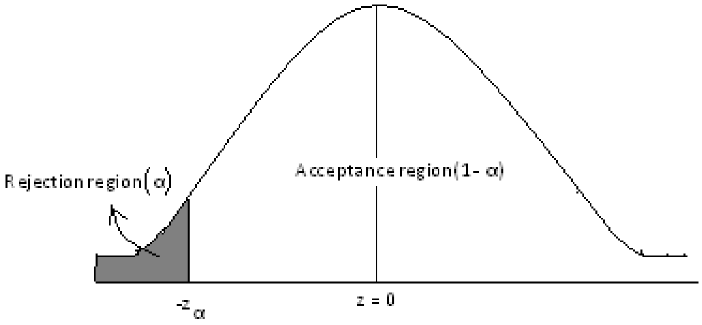
(a) Right tailed test:

In case of right tailed test $P\left(z > z_{\alpha}\right) = \alpha$



(b) Left tailed test:

In case of right tailed test $P\left(z < -z_{\alpha}\right) = \alpha$



Procedure for testing of Hypothesis:

The working rule or procedure may be adopted in testing of a hypothesis:

1) **Null Hypothesis** : Construct null hypothesis H_0.

2) **Alternative Hypothesis** : Construct alternative hypothesis H_1 and decide whether it is one -tailed or two-tailed test.

3) **Level of Significance** : Select the appropriate level of significance(α) in advance.

4) **Test Statistic** : Compute the test statistic z under the null hypothesis $z = \frac{\bar{X} - E[X]}{S.E[X]}$

5) **Conclusion** : Compare the calculated z value with tabulated value at the given level of significance α

If $|z| < z_\alpha$, Accept the null hypothesis H_0 at α level of significance

If $|z| > z_\alpha$, Reject the null hypothesis H_0 and accept alternative hypothesis H_1at α level.

Critical value z_α	**Level of Significance**		
	1% (.01)	**5%(.05)**	**10% (.1)**
Two-tailed test	$\lvert z_{\alpha/2}\rvert$=2.58	$\lvert z_{\alpha/2}\rvert$=1.96	$\lvert z_{\alpha/2}\rvert$=1.64
One-tailed test	z_α=2.33	z_α=1.64	z_α=1.28
Left-tailed test	-z_α=2.33	-z_α=1.64	-z_α=1.28

Test of Significance of single mean-Large Samples

Let a random sample of size n ($n \geq 30$) has the sample mean $\overline{x}$, and μ be the population mean. Also the population mean μ has a specified value μ_0.

Null Hypothesis is H_0: $\overline{x} = \mu$ ($\mu = \mu_0$) i.e , "there is no significance difference between the sample mean and population mean " or the sample has been drawn from the parent population.

Alternative Hypothesis is

- $H_1 : \overline{x} \neq \mu (\mu \neq \mu_0)$
- $H_1 : \overline{x} > \mu (\mu > \mu_0)$
- $H_1 : \overline{x} < \mu (\mu < \mu_0)$

Test statistic: when the standard deviation σ of population is known.

In this case, standard Error of mean, S.E ($\overline{x}$) = $\frac{\sigma}{\sqrt{n}}$, where n=sample size, σ = S.D of population.

The test statistic is $z = \frac{x-\mu}{S.E\left(\overline{x}\right)} = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$, where μ the sample is mean.

Example 1: According to the norms established for a mechanical aptitude test, persons who are 18 years old have an average height of 73.2 with a standard deviation of 8.6.If 40 randomly selected persons of that age averaged 76.7, test the hypothesis μ =73.2 against the alternative hypothesis $\mu > 73.2$ at the 0.01 level of significance.

Solution: Given n=40, μ =73.2, $\overline{x}$ =mean of the sample=76.7 and σ =S.D of population=8.6

Null Hypothesis $H_0 : \mu$ =73.2

Alternative Hypothesis $H_1 : \mu > 73.2$

Level of significance: α =0.01

Test statistic is $z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{76.7-73.2}{\frac{8.6}{\sqrt{40}}} = 2.57$

Conclusion: The tabulated value of $z_\alpha = 2.33$ at 1% l.o.s

Since $|z_{cal}| > |z_\alpha|$, Reject the Null Hypothesis H_0 at 1% level of significance

Example 2: A sample of 64 students have a mean weight of 70 kg .Can this be regarded as a sample from a population with mean weight 56 kg and standard deviation 25 kg.

Solution:

Given $\overline{x}$ = mean of the sample =70 kg, μ =mean of the population =56 kg

σ =S.D of population =25 kg and n=sample size =64

Null Hypothesis H_0 : sample is regarded to be taken from the population.

Alternative Hypothesis H_1: sample cannot be regarded as taken coming from the population.

Level of significance: α =0.05(assumption)

The test statistic is $z = \dfrac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \dfrac{70-56}{\left(\frac{25}{\sqrt{64}}\right)} = 4.48$

Conclusion: The tabulated value of $z_{\alpha/2} = 1.96$ at 5% l.o.s

Since $|z_{cal}| > |z_{\alpha/2}|$, Reject Null Hypothesis H_0 at 5% level of significance

Example 3: In a random sample of 60 workers, the average time taken by them to get to work is 33.8 minutes with a standard deviation of 6.1 minutes. Can be reject the null hypothesis $\mu = 32.6$ minutes in favor of alternative null hypothesis $\mu > 32.6$ at α =0.025 level of significance.

Solution: Given n=60, $\bar{x}$ =33.8, μ =32.6 and σ =6.1

Null hypothesis H_0 : μ =32.6

Alternative Hypothesis H_1 : μ >32.6 (right tailed)

Level of significance: α =0.025

The test statistic is $z = \dfrac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \dfrac{33.8-32.6}{\left(\frac{6.1}{\sqrt{60}}\right)} = \dfrac{1.2}{0.7875} = 1.5238$

Conclusion: The tabulated value of $z_{\alpha} = 1.96$ at 0.025% l.o.s

Since $|z_{cal}| < |z_{\alpha}|$, Accept Null Hypothesis H_0 at 0.025% level of significance

Example 4: An ambulance service claims that it takes on the average less than 10 minutes to reach its destination in emergency calls. A sample of 36 calls has a mean of 11 minutes and the variance of 16 minutes. Test the claim at 0.05 level significance.

Solution: Given n=36, $\bar{x} = 11, \mu = 10$ and $\sigma = \sqrt{16} = 4$

Null hypothesis $H_0 : \mu = 10$

Alternative Hypothesis $H_1 : \mu < 10$

Level of significance: $\alpha = 0.05$

The test statistic is, $z = \dfrac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \dfrac{11-10}{4/\sqrt{36}} = \dfrac{6}{4} = 1.5$

Conclusion: The tabulated value of $z_{\alpha} = 1.645$ at 5% l.o.s

Since $|z_{cal}| < |z_{\alpha}|$, Accept Null Hypothesis H_0 at 5% level of significance

Example 5: A sample of 400 items is taken from a population whose standard deviation is 10. The mean of the sample is 40. Test whether the sample has come from a population with mean 38. Also calculate 95% confidence interval for the population.

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

Null hypothesis $H_0: \ \mu = 38$

Alternative Hypothesis $H_1: \ \mu \neq 38$

Level of significance: α =0.05

The test statistic is: $z = \dfrac{\bar{x}-\mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{40-38}{\left(\dfrac{10}{\sqrt{400}}\right)} = 4$

Conclusion: The tabulated value of $z_{\alpha/2}$ =1.96 at 5% l.o.s

Since $|z_{cal}| > |z_{\alpha/2}|$, Reject Null Hypothesis H_0 at 5% level of significance

i.e the sample is not from the population whose mean is 38.

95% confidence interval is $\left(\bar{x}-1.96\dfrac{\sigma}{\sqrt{n}}, \bar{x}+1.96\dfrac{\sigma}{\sqrt{n}}\right)$ = (39.02, 40.98)

Exercise:

1. Forty bearings made by a certain process have a mean diameter of 0.5060cm with a standard deviation of 0.0040cm. Assuming that the data may be taken as a random sample from a normal distribution, Construct a 99% confidence interval for the actual average diameter of the bearings?

2. A random sample of size 225 is taken whose mean is 80. Can this be regarded as a sample from a population with mean 82 and S.D15 at 0.05 l.o.s. [Ans: Reject H_0]

3. A oceanographer wants to check whether the depth of the ocean in a certain region is 57.4 fathoms, as had previously been recorded. If recordings taken at 40 random locations in the given region yielded a sample mean of 59.1 fathoms with S.D 5.2 fathoms. What can be concluded at the 5% l.o.s, to reject null hypothesis that the mean depth is 57.4 fathoms? [Ans: Reject H_0]

4. The mean life time of a sample of 100 light tubes produced by a company is found to be 1560hrs with a population S.D of 90 hrs. Test the hypothesis for 0.05 l.o.s that the mean life time of the tubes produced by the company is 1580 hrs. [Ans: Reject H_0]

5. It is claimed that a random sample of 49 tires has a mean life of 15200 km. This sample was drawn from a population whose mean is 15150kms and a standard deviation of 1200 km. Test the significance at 0.05 levels. [Ans: Accept H_0]

Test of Significance of difference of means-Large Samples

Let $\overline{x_1}$ and $\overline{x_2}$ be the sample means of two independent large random samples sizes n_1 and n_2 drawn from two populations having means μ_1 and μ_2 and standard deviations σ_1 and σ_2

To test whether the two-population means are equal:

Null Hypothesis be $H_0 : \mu_1 = \mu_2$

Alternative Hypothesis is $H_1 = \mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

Test statistic is $$Z=\frac{\left(\overline{x_1}-\overline{x_2}\right)-\delta}{S.Eof\left(\overline{x_1}-\overline{x_2}\right)}=\frac{\left(\overline{x_1}-\overline{x_2}\right)-\delta}{\sqrt{\frac{\sigma^2{}_1}{n_1}+\frac{\sigma^2{}_2}{n_2}}}$$

Where $\delta = \mu_1 - \mu_2$ (=given constant)

If $\delta \neq 0$, the two populations are different.

If $\delta = 0$, the two populations have same means.

Under $H_0 : \mu_1 = \mu_{2,}$, the test statistic becomes $z=\frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{\sigma^2{}_1}{n_1}+\frac{\sigma^2{}_2}{n_2}}}$

Note: If the samples have been drawn from the population with common S.D. σ, then

$\sigma^2{}_1 = \sigma^2{}_2 = \sigma^2$ Hence $z=\frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}}}$

If σ is not known we can use an estimate of σ^2 given by $\sigma^2=\frac{n_1 s^2{}_1+n_2 s^2{}_2}{n_1+n_2}$

Example 1: The means of two large samples of sizes 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of S.D 2.5 inches.

Solution: Let μ_1 and μ_2 be the means of two populations.

Given n_1 =1000, n_2 =2000 and $\overline{x_1}$ =67.5 inches, $\overline{x_2}$ =68 inches, $\sigma = 2.5$ Inches

Null Hypothesis H_0: $\mu_1 = \mu_2$

Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$

Level of Significance : $\alpha = 0.05$

The test statistic is, $z=\frac{\overline{x_1}-\overline{x_2}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}=\frac{67.5-68-0}{\sqrt{\frac{1}{1000}+\frac{1}{2000}}}$ = -5.16

Conclusion: The tabulated value of $z_{\alpha/2}$ =1.96 at 5% l.o.s

Since $|z_{cal}| > |z_{\alpha/2}|$, Reject Null Hypothesis H_0 at 5% level of significance

we conclude that the samples are not drawn from the same population of S.D. 2.5 inches.

Example 2: A researcher wants to know the intelligence of students in a school .He selected two groups of students. In the first group there 150 students having mean IQ of 75 with a S.D of 15 in the second group there are 250 students having mean IQ of 70 with S.D. of 20. Is there a significant difference between the intelligence of two groups at 1% lo.s..

Solution: Given n_1=150, $\overline{x_1}$=75, σ_1=15

And n_2=250, $\overline{x_2}$=70, σ_2=20

Null Hypothesis H_0 : No significant difference between the intelligence of two groups

$$\mu_1 = \mu_2$$

Alternative Hypothesis is $H_1 : \mu_1 \neq \mu_2$

Level of Significance: $\alpha = 0.01$

The test statistic is , $z = \dfrac{\overline{x_1} - \overline{x_2} - \delta}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \dfrac{75-70-0}{\sqrt{\dfrac{15^2}{150} + \dfrac{20^2}{250}}} = \dfrac{5\sqrt{5}}{\sqrt{17}} = 2.7116$

Conclusion: The tabulated value of $z_{\alpha/2} = 2.33$ at 1% l.o.s

Since $|z_{cal}| > |z_{\alpha/2}|$, Reject Null Hypothesis H_0 at 1% level of significance

Conclude that the groups have not been taken from the same population.

Example 3: The mean life of a sample of 10 electric bulbs (or motors) was found to be 1456 hours with S.D of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 hours with S.D of 398 hours. Is there a significant difference between the means of two batches?

Solution: It is given that

n_1 =Sample size of first batch =10 n_2 =Sample size of second batch=17

$\overline{x_1}$ =Mean life of first batch =1456 $\overline{x_2}$ =Mean life of second batch=1280

σ_1 =Standard deviation of first batch=423 σ_2 =Standard deviation of second batch=398

Null Hypothesis $H_0 : \mu_1 = \mu_2$

Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$

Level of significance: $\alpha = 0.05$

The test statistic is $z = \dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\dfrac{\sigma^2{}_1}{n_1} + \dfrac{\sigma^2{}_2}{n_2}}} = \dfrac{176}{164.96} = 1.067$

Conclusion: The tabulated value of $z_{\alpha/2} = 1.96$ at 5% l.o.s

Since $|z_{cal}| < |z_{\alpha/2}|$, Accept Null Hypothesis H_0 at 5% level of significance

There is no difference between the mean life of electric bulbs of two batches.

Example 4: The average marks scored by 32 boys are 72 with a S.D of 8. While that for 36 girls is 70 with a S.D of 6. Does this indicates that the boys perform better than girls at level of significance 0.05?

Solution: Here $\bar{x}$ =72, $\bar{y}$ =70, σ_1 =8, σ_2 =6, n_1 =32, n_2 =36

Null Hypothesis be $H_0 : \mu_1 = \mu_2$

Alternative Hypothesis is $H_1 : \mu_1 > \mu_2$

Level of significance: $\alpha = 0.05$

Test statistic $z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ =1.1547

Conclusion: The tabulated value of $z_{\alpha/2} = 1.96$ at 5% l.o.s

Since $|z_{cal}| < |z_{\alpha/2}|$, Accept Null Hypothesis H_0 at 5% level of significance

Conclude that the performance of boys and girls is the same.

Example 5: A company claims that its bulbs are superior to those of its main competitor. If a study showed that a sample of 40 0f its bulbs have a mean life time of 647 hrs of continuous use with a S.D of 27 hrs .While a sample of 40 bulbs made by its main competitor had a mean life time of 638 hrs of continuous use with a S.D of 31 hrs. Test the significance betweenthe difference of two means at 5% level.

Solution: Here $\bar{x}$ =647, $\bar{y}$ =638, σ_1 =27, σ_2 =31, $n_1 = n_2$ =40

Null Hypothesis be $H_0 : \mu_1 = \mu_2$

Alternative Hypothesis is $H_1 : \mu_1 > \mu_2$

Level of significance: $\alpha = 0.05$

Test statistic $z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ =1.38

Conclusion: The tabulated value of $z_\alpha = 1.645$ at 5% l.o.s

Since $|z_{cal}| < |z_\alpha|$, Accept Null Hypothesis H_0 at 5% level of significance

The difference between the two sample means is not significant.

Exercise:

1. In a certain factory there are two independent processes for manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 gms with a S.D of 12 gms. While the corresponding figures in sample of 400 items from the other process are 124 and 14 gms. Test the significance between the difference of two means at 5% level. Construct Confidence interval. [Ans: Reject H_0, (-6.02, -1.976)]

2. Two types of new cars produced are tested for petrol mileage, one sample is consisting of 42 cars averaged 15kmpl, while other sample is consisting of 80 cars averaged 11.5 kmpl with population variances 2.0 and 1.5 respectively. Test whether there is any significant difference in the petrol consumption of these two types of cars at 0.01 level. [Ans: Reject H_0]

3. At a certain large university a sociologist speculates that male students spend considerably more money on junk food than do female students. To test her hypothesis, the sociologist randomly selects from the register's records the names of 200 students. Of these, 125 are men and 75 are women. The sample mean of the average amount spent on junk food per week by the men is 400 and S.D is 100. For the women the sample mean is 450 and S.D is 150. Test the difference between the means at 0.05 level [Ans: Reject H_0]

4. A company claims that alloying reduces resistance of electric wire by more than 0.05 ohms. To test this claim samples of standard wire and alloyed wire are tested yielding the following :

Type of wire	Sample size	Mean resistance (ohms)	Standard Deviation (ohms)
Standard	32	0.136	0.004
Alloyed	32	0.083	0.005

Can the claim be substantiated at 0.05 l.o.s. [Ans: Reject H_0]

Test of Significance for Single Proportion-Large Samples

Suppose a large random sample of size n has a sample proportion p of members possessing a certain attribute (i.e, proportion of successes).To test the hypothesis that the proportion P in the population has a specified value P_0.

Null Hypothesis: H_0 : P = P_0 (P_0 is a particular value of P)

Alternative Hypothesis:

(i) $H_1 : P \neq P_0$ or (ii) $H_1 : P > P_0$ or (iii) $H_1 : P < P_0$

Since n is large, the sampling distribution of p is approximately normal.

If H_0 is true, the test statistic is $z = \dfrac{p - P_0}{S.Eofp} \Rightarrow z = \dfrac{p-P}{\sqrt{\dfrac{PQ}{n}}}$

where p, the sample proportion is approximately normally distributed.

Example 1: A manufacturer claimed that at least 95% of the equipment which he supplied to a factory conformed to specifications. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. Test his claim at 5% level of significance.

Solution: Given sample size, n=200

Number of pieces confirming to specification =200-18=182

p=Proportion of pieces confirming to specifications $= \dfrac{182}{200} = 0.91$

P=Population proportion $= \dfrac{95}{100} = 0.95$

Null Hypothesis H_0 : The proportion of pieces confirming to specifications i.e P=95%

Alternative Hypothesis H_1 : P<0.95 (left-tailed test)

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

Test statistic: $z = \dfrac{p-P}{\sqrt{\dfrac{PQ}{n}}} = \dfrac{0.91-0.95}{\sqrt{\dfrac{0.95\times 0.05}{200}}} = \dfrac{-0.04}{0.0154} = -2.59$

Conclusion: The tabulated value of $z_\alpha = 1.645$ at 5% l.o.s

Since $|z_{cal}| > |z_\alpha|$, Reject the Null Hypothesis H_0 at 5% level of significance and conclude that the manufactures claim is rejected.

Example 2: In a sample of 1000 people in Karnataka 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance.

Solution:

Given n=1000

P=sample proportion of rice eaters = $\dfrac{540}{1000} = 0.54$

P=population proportion of rice eaters= $\dfrac{1}{2} = 0.5$ and Q=0.5

Null Hypothesis H_0 : both rice and wheat are equally popular in the state.

Alternative Hypothesis H_1 : $P \neq 0.5$ (two-tailed alternative)

Level of significance: $\alpha = 1\%$

Test statistic is $z = \dfrac{p-P}{\sqrt{\dfrac{PQ}{n}}} = \dfrac{0.54-0.5}{\sqrt{\dfrac{0.5\times 0.5}{1000}}} = 2.532$

Conclusion: The tabulated value of $z_{\alpha/2} = 2.58$ at 1% l.o.s

Since $|z_{cal}| < |z_{\alpha/2}|$, Accept H_0 at 1% level of significance and conclude that both rice and wheat are equally popular in the state.

Example 3: If 80 patients are treated with an antibiotic 59 got cured. Find a 99% confidence limits to the true population of cure.

Solution: n=80, x=59 and $P = \dfrac{x}{n} = \dfrac{59}{80} = 0.7375$, Q=1-P=1-0.7375=0.2625 , $z_{\frac{\alpha}{2}} = 2.58$

Confidence interval is $\left(P - z_{\alpha/2}\sqrt{\dfrac{PQ}{n}}, P + z_{\alpha/2}\sqrt{\dfrac{PQ}{n}}\right)$ = (0.59, 0.88)

Exercise:

1. In a random sample of 125 soft drinkers, 68 prefer Thums up to pepsi. Test the null hypothesis P=0.5 against alternative hypothesis P>0.5. [Ans: Accept H_0]

2. A manufacturer of bulbs claimed that the % defective in his product does not exceed 6. A sample of 40 bulbs is found to contain 5 defective. Would you consider the claim justified?

3. In a study designed to investigate whether certain detonators used with explosives in coal mining meet the requirement that at least 90% will ignite the explosive when charged. It is found that 174 of 200 detonators function properly. Test the null hypothesis $P = 0.90$ against the alternative hypothesis $P \neq 0.90$ at the 0.05 level of significance. [Ans: Accept H$_0$]

4. In a sample of 500 from a village in Rajasthan, 280 are found to be wheat eaters and the rest are rice eaters. Can we assume that the both articles are equally popular. [Ans: Reject H$_0$]

5. A random sample of 500 pineapples was taken from a large consignment and 65 are found to be bad. Find the percentage of bad pineapples in the consignment. [Ans: (0.085, 0.175)]

6. In a hospital 480 females and 520 males were born in a week. Do these figures confirm the hypothesis that males and females are born in equal number? [Ans: Accept H$_0$]

Test of Significance for difference between two sample Proportions-Large Samples

Let p_1 and p_2 be the sample proportions in two large random samples of sizes n_1 and n_2 drawn from two populations having proportions P_1 and P_2

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

Null hypothesis $H_0 : P_1 = P_2$

Alternative hypothesis $H_1 : P_1 \neq P_2$

Test statistic: There are two ways of computing a test statistic z.

(a)When the population proportions P_1 and P_2 are known.

In this case, $Q_1 = 1 - P_1$ and $Q_2 = 1 - P_2$ and p_1, p_2 are sample proportions.

Standard Error of Difference = S.E ($p_1 - p_2$) = $\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

Hence the test statistic is $z = \frac{p_1 - p_2}{S.E(p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$

(b) When the population proportions P_1 and P_2 are not known but sample proportions p_1 and p_2 are known. In this case we have two methods to estimate P_1 and P_2.

(i) METHOD OF SUBSTITUTION: In this method sample proportions p_1 and p_2 are substituted for P_1 and P_2.

S.E (p_1 - p_2)= $\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$ Hence the test statistic is Z= $\frac{p - p_2}{S.E(p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}$

(ii)METHOD OF POOLING: In this method, the estimated value for the two population proportions is obtained by pooling the two sample proportions p_1 and p_2 into a single proportion p by the formula given below.

Sample proportion of two samples or estimated value of p is given by

P=$\dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$, so that $q = 1 - p$

Hence the test statistic is Z=$\dfrac{p_1 - p_2}{S.E(p_1 - p_2)} = \dfrac{p_1 - p_2}{\sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

Example 1: Random sample of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favor of the proposal. Test the hypothesis that proportions of men and women in favor of the proposal are same, at 5% level.

Solution: Given sample sizes, n_1=400, n_2=600

Proportion of men, $p_1 = \dfrac{200}{400} = 0.5$

Proportion of women, $p_2 = \dfrac{325}{600} = 0.541$

p = $\dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ = 0.525 , q=1-p = 0.475

Null hypothesis: Assume that there is no significant difference between the option of men and women as far as proposal of flyover is concerned.

$$H_0 : p_1 = p_2 = p$$

Alternative hypothesis $H_1 : p_1 \neq p_2$ (two tailed)

Test statistic is $z = \dfrac{p_1 - p_2}{\sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ [method of pooling]

$$z = \frac{0.5 - 0.541}{\sqrt{0.525 \times 0.425\left(\dfrac{1}{400} + \dfrac{1}{600}\right)}} = \frac{-0.041}{0.032} = -1.28$$

Conclusion: The tabulated value of $z_{\alpha/2} = 1.96$ at 5% l.o.s

Since $|z_{cal}| < |z_{\alpha/2}|$, Accept the Null Hypothesis H_0 at 5% level of significance

i.e., there is no difference of opinion between men and a woman as far as proposal of flyover is concerned.

Example 2: In two large populations, there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

Solution: Given $n_1 = 1200, n_2 = 900$

P_1 = Proportion of fair haired people in the first population= $\frac{30}{100} = 0.3$

P_2 = Proportion of fair haired people in the second population= $\frac{25}{100} = 0.25$

Null hypothesis: Assume that the sample proportions are equal i.e ., the difference in population proportions is likely to be hidden in sampling i.e., $H_0 : p_1 = p_2 = p$

Alternative hypothesis $H_1 : p_1 \neq p_2$ (two tailed)

Test statistic is, $z = \dfrac{P_1 - P_2}{\sqrt{\dfrac{P_1Q_1}{n_1} + \dfrac{P_2Q_2}{n_2}}}$

Where $Q_1 = 1 - P_1 = 1 - 0.3 = 0.7$ $\quad Q_2 = 1 - P_2 = 1 - 0.25 = 0.75$

$$z = \frac{0.3 - 0.25}{\sqrt{\dfrac{0.3 \times 0.7}{1200} + \dfrac{0.25 \times 0.75}{900}}} = \frac{0.05}{0.0195} = 2.56$$

Conclusion: The tabulated value of $z_{\alpha/2} = 1.96$ at 5% l.o.s

Since $|z_{cal}| > |z_{\alpha/2}|$, reject the Null Hypothesis H_0 at 5% level of significance

i.e, the sample proportions are not equal. Thus we conclude that the difference in population proportions is unlikely that the real difference will be hidden.

Exercise:

1.100 articles from a factory are examined and 10 are found to be defective. 500 similar articles from a second factory are found to be 15 defective. Test the significance between the difference of two proportions at 5% level.

2. Before an increase on excise duty on tea 500 people out of 900 found to have the habit of having tea. After an increase on excise duty 250 have the habit of having tea among 1100. Is there any decrease in the consumption of tea. Test at 5% level. [Ans: Reject H_0]

3. On the basis of their total scores, 200 candidates of a civil service examination are divided into two groups, the upper 30% and remaining 70%. Consider the first question examination. Among the first group, 40 had the correct answer, whereas among the second group, 80 had the correct answer. On the basis of these result, can one conclude that the first question is not good at discriminating ability of the type being examined here? [Ans: Accept H_0]

4. A study shows that 16 of 200 tractors produced on one assembly line required extensive adjustments before they could be slipped. While the same was true for 14 of 400 tractors produced on another assembly line. At 1% l.o.s, does this support the claim that the second production line does the superior work? [Ans: Reject H_0]

5. In 2 large populations, there are 30% and 25% respectively of blue-eyed people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations? [Ans: Reject H_0]

6. In a year there are 956 births in a town A, of which 52.5% were males, while in towns A and B combined, this proportion in a total of 1406 births was 0.496. Is there any significant difference in the proportions of male births in the two towns? [Ans: Reject H_0]

COMPUTER ORIENTED STATISTICAL METHODS(A8005)
#5. Testing of Hypothesis for Small Samples

Small sample:

If the size of the sample $(n) < 30$, the sample is said to be small sample.

when the sample is small(n<30), we can use normal distribution to test for a specified population mean or difference of two population means as in large sample tests only when the sample is drawn from a normal population whose S.D., σ is known . If σ is not known , we cannot proceed as above. If a population is normally distributed the Sampling distribution of the sample mean for any sample size is also normally distributed whether σ is known or not.

Degrees of freedom: (d.f)

In general , the number of degrees of freedom is equal to the total number of observations less the number of independent constraints imposed on the observations. For example in a set of data of n observations, if k is the number of independent constraints then $\upsilon = n - k$.

.

Test of significance for small samples:.

A very important aspect of the sampling theory is the study of tests of significance, which Enable us to decide on the basis of the sample results, if

(i) The deviation between the observed sample statistic and the hypothetical parameter value is significant.

(ii) The deviation between two sample statistics is significant.

The following are some important tests for small samples:

(i) Students $'t'$ test

(ii) F -test

(iii) χ^2 - test

Student's 't' Test:

It is used for testing of hypothesis when the sample size is small and population S.D ,σ is not known.

Let $\overline{x}$ = Mean of a sample

n = Size of the sample

σ = Standard deviation of the sample

μ = Mean of the population supposed to be normal

Then the student's t is defined by the statistic

(i) $t = \dfrac{\overline{x} - \mu}{s/\sqrt{n-1}}$ (if s is given)

(ii) $t = \dfrac{\overline{x} - \mu}{S/\sqrt{n}}$ where $S^2 = \dfrac{\sum_i (x_i - \overline{x})^2}{n-1}$ (if data is given)

S^2 is called the unbiased estimate of population variance σ²

Confidence or Fiducial limits for μ :

Suppose we want to find the sample data the limits within which the population mean will lie with a probability of 0.95. The limits are called the 95% confidence limits of the population mean for the given sample.

If $t_{0.05}$ is the table value of t for $(n-1)$ degrees of freedom at 5% level of significance , then 95% confidence limits for μ are given by $\bar{x} \pm t_{0.05} \cdot \frac{S}{\sqrt{n}}$

For $P(|t| > t_{0.05}) = 0.05 \; i.e., P(|t| \le t_{0.05}) = 0.95$

95% confidence limits for μ are given by

$$|t| \le t_{0.05} \quad \text{i.e.,} \quad \left|\frac{\bar{x}-\mu}{S/\sqrt{n}}\right| \le t_{0.05} \quad \text{or} \quad \frac{|\bar{x}-\mu|}{S/\sqrt{n}} \le t_{0.05}$$

$$\Rightarrow -t_{0.05} \le \frac{\bar{x}-\mu}{S/\sqrt{n}} \le t_{0.05}$$

$$\Rightarrow \bar{x} - t_{0.05}.S/\sqrt{n} \le \mu \le \bar{x} + t_{0.05}.S/\sqrt{n}$$

Similarly 99% confidence limits for μ are $\bar{x} \pm t_{0.05}.S/\sqrt{n}$

Where $t_{0.01}$ is the tabulated value of t for $(n-1)$ degrees of freedom at 1% level of significance.

Example 1: The average breaking strength of the steel rods is specified to be 18.5 thousand pounds .To test this sample of 14 rods were tested .The mean and standard deviations obtained were 17.85 and 1.955 respectively .Is the result of experiment significant?

Solution:

Given sample size , n = 14

Sample mean $\bar{x}$ = 17.85

S.D (s) = 1.955

Population mean , μ = 18.

Degrees of freedom = $n-1$ = 13

1. Null Hypothesis H_0 : The result of the experiment is not significant .
2. Alternative Hypothesis H_1 : $\mu \ne 18.5$
3. Level of significance : α = 0.05
4. The test statistic is , $t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}} = \frac{17.85-18.85}{1.955/\sqrt{13}} = \frac{\overset{1}{0.65}}{0.542} = -1.199$

$$\therefore |t| = 1.199$$

i.e., Calculated t = 1.199

Tabulated t at 5% level of significance for 13 d.f for two tailed test = 2.16

Since calculated $t <$ tabulated t, we accept the null Hypothesis H_0 at 5% level and conclude that the result of the experiment is not significant.

Example 2: A random sample of 10 boys had the following I.Q's: 70, 120, 110, 101, 88, 83, 95, 98, 107 and 100.
(a) Do these data support the assumption of a population mean I.Q of 100?
(b) Find a reasonable range in which most of the mean I.Q values of samples of 100 boys lie.

Solution: (a) Here S.D and mean of the sample is not given directly. We have to determine these S.D and mean as follows.

Mean , $\overline{x} = \dfrac{\sum x}{n} = \dfrac{972}{10} = 97.2$

x	$x-\overline{x}$	$\left(x-\overline{x}\right)^2$
70	-27.2	739.8
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
88	-9.2	84.44
83	-14.2	201.64
95	-2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84
972		**1833.60**

We know that $S^2 = \dfrac{1}{n-1}\sum_i (x_i - \overline{x})^2 = \dfrac{1833.60}{9}$

Standard deviation , $S = \sqrt{203.73} = 14.27$

1. Null hypothesis is H_0: The data support the assumption of a population mean I.Q of 100 in the population.
2. Alternative Hypothesis H_1: $\mu \neq 100$
3. Level of significance: $\alpha = 0.05$
4. The Test Statistic is $t = \dfrac{\overline{x}-\mu}{S/\sqrt{n}} = \dfrac{97.2-100}{14.27/\sqrt{10}} = -0.62$

$\therefore$ $|t| = 0.62$ i.e., Calculated value of t = 0.62

Tabulated value of t for (10-1) d.f i.e., 9 d.f at 5% level of significance is 2.26(two – tailed list).

Since calculated value of $t <$ tabulated value of t , we accept the null hypothesis H_0 .i.e., the data support the assumption of mean I.Q of 100 in the population.

(b) The 95% confidence limits are given by $\overline{x} \pm t_{0.05}.S/\sqrt{n}$

= 97.2 $\pm$ 2.26 x 4.512 = 97.2 $\pm$ 10.198 = 107.4 and 87

$\therefore$ The 95% confidence limits within which the mean I.Q values of samples of 10 boys will lie is (87 , 107.40)

Exercise:

1. A mechanist is making engine parts with axle diameters of 0.700 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a standard deviation of 0.040 inch. Compute the statistic you would use to test whether the work is meeting the specifications.
   [Ans: 3.15, reject H_0]

2. The mean weekly sales of soap bars is departmental stores was 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful?
   [Ans: 1.97, reject H_0]

3. The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches? Test at 5% significance level assuming that for 9 degrees of freedom $P(t > 1.83)$ =0.05.

4. Given a random sample of 5 pints from different production lots. Test whether the fat content of a certain kind of ice cream exceeds 14%. What can we conclude at 0.01 l.o.s about the null hypothesis $\mu = 14\%$ if the sample has mean $\overline{x} = 14.9\%$ and S.D $s = 4.2\%$.

5. A random sample from a company's extensive files shows that the order of a certain kind of machinery were filled respectively in 10, 12, 19, 14, 15, 18, 11 and 13 days. Use 0.01 l.o.s to test the claim that on the average such orders are filled in 10.5 days. Assume normality?
   [Ans: 3.087, Accept H_0]

t-test for difference of means:

Let $\overline{x}$ and $\overline{y}$ be the means of two independent samples of sizes n_1 and n_2 ($n_1 < 30$, $n_2 < 30$) drawn from two normal population having means μ_1 and μ_2. To test whether the two population means are equal (i.e., to test whether the difference $\mu_1 - \mu_2$ is significant).

Let the Null Hypothesis be H_0: $\mu_1 = \mu_2$
Then the Alternative Hypothesis is H_1: $\mu_1 \neq \mu_2$

Test statistic t given by $t = \dfrac{\overline{x} - \overline{y}}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ follows t – distribution with ($n_1 + n_2 - 2$) d.o.f.

If $\sigma_1 = \sigma_2 = \sigma$, then an unbiased estimate S^2 of the common variance σ^2 is given by

$$S^2 = \frac{1}{n_1 + n_2 - 2}[\sum(x_i - \overline{x})^2 + \sum(y_i - \overline{y})^2] \text{ . Here } \overline{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i \text{ , } \overline{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i \text{ and}$$

Note: If S.D's s_1 and s_2 are given, $S^2 = \dfrac{n_1 s_1{}^2 + n_2 s_2{}^2}{n_1 + n_2 - 2}$

We calculate the value of | t | and compare this value with the table value at α level of significance. If the calculated value of | t | > the table value , we reject H₀ at α level. Otherwise we accept H₀.

Confidence limits for the difference of two population means are $(\bar{x}-\bar{y}) \pm t_\alpha . S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

Example 1: A group of 5 patients treated with medicine A weigh 42, 39, 48, 60 and 41 kgs . Second group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 69 and 62 kgs. Do you agree with the claim that medicine B increases the weight significantly.

Solution: Calculation for sample means and S.D's

x	$(x-\bar{x})$	$(x-\bar{x})^2$	y	$(y-\bar{y})$	$(y-\bar{y})^2$
42	-4	16	38	-19	361
39	-7	49	42	-15	225
48	2	56	56	-1	1
60	14	64	64	7	49
41	-5	68	68	11	21
		69	69	12	44
		62	62	5	25
230	0	399	399	0	926

Now $\bar{x} = \frac{230}{5} = 46$, $\bar{y} = \frac{399}{7} = 57$

And $\sum_{i=1}^{5}(x_i-\bar{x})^2 = 290$, $\sum_{j=1}^{7}(y_i-\bar{y})^2 = 926$

$$\therefore S^2 = \frac{1}{n_1+n_2-2}[\sum(x_i-\bar{x})^2 + \sum(y_i-\bar{y})^2] = \frac{1}{5+7-2}[290+926] = 121.6$$

$$\Rightarrow S = 11.03$$

1. Null hypothesis is H₀: There is no significant difference between the medicines A and B as regards their effect on increase in weight i.e., H₀: $\mu_1 = \mu_2$
2. Alternative Hypothesis H₁: $\mu_1 > \mu_2$
3. Level of significance: α = 0.05
4. The Test Statistic is $t = \frac{\bar{x}-\bar{y}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{46-57}{(11.03)\sqrt{\frac{1}{5}+\frac{1}{7}}} = \frac{-11}{6.46} = -1.7$

$\therefore$ Calculated value of | t | =1.7
Tabulated value of t for 5+7-2 = 10 d.f at 5% level of significance is 1.81.
Since calculated value of t < tabulated value of t ,we accept the null hypothesis H₀.i.e The medicines A and B do not differ significantly as regards their effect on increase in weight.

Example 2: To examine the hypothesis that the husbands are more intelligent than the wives, an investigator took a sample of 10 couples and administered them a test which measures the I.Q. The results are as follows:

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

Test the hypothesis with a reasonable test at the level of significance of 0.05

Solution: We have $n_1 = 10$ and $n_2 = 10$ and

$$\bar{x} = \frac{1}{10}[117+105+97+105+123+109+86+78+103+107] \quad = \frac{1}{10}(1030) = 103$$

$$\bar{y} = \frac{1}{10}[106+98+87+104+116+95+90+69+108+85] \quad = \frac{1}{10}(958) = 95.8$$

X	$x-\bar{x}$	$(x-\bar{x})^2$	y	$y-\bar{y}$	$(y-\bar{y})^2$
117	14	196	106	10.2	104.04
105	2	4	98	2.2	4.84
97	-6	36	87	-8.8	77.44
105	2	4	104	8.2	67.24
123	20	400	116	20.2	408.04
109	6	36	95	-0.8	0.64
86	-17	289	90	-5.8	33.64
78	-25	625	69	-26.8	718.24
103	0	0	108	12.2	148.84
107	4	16	85	-10.8	116.64
1030		**1606**	**958**		**1679.6**

$$S^2 = \frac{1}{n_1+n_2-2}[\sum(x_i-\bar{x})^2 + \sum(y_i-\bar{y})^2] = \frac{1}{18}[1606+1679.6] = \frac{1}{18}(3285.6) = 182.53$$

$$\therefore S = 13.51$$

1. Null hypothesis is H₀: $\mu_1 = \mu_2$ (i.e., no difference in I.Q)
2. Alternative Hypothesis H₁: $\mu_1 > \mu_2$ (i.e., husbands are more intelligent than wives)
3. Level of significance: α = 0.05
4. The Test Statistic is $t = \dfrac{\bar{x}-\bar{y}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \dfrac{103-95.8}{(13.51)\sqrt{\frac{1}{10}+\frac{1}{10}}} = 1.19168$

Since calculated t = 1.19168 < tabulated t = 1.734 , we accept the null hypothesis H_0 i.e., There is no difference in I.Q's.

Example 3: Find the S.E of difference between the means and also find the confidence interval for the difference of means at 0.05 level for the following data.

Sample No.	Size	Means	S.D
I	9	69	4
II	10	68	5

Solution: $S=\sqrt{\frac{n_1 s_1^2+n_2 s_2^2}{n_1+n_2}}=4.81$ and $S.E=S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}=2.21$. $t_{\alpha/2}=2.11$ at 17 d.o.f

Confidence interval is $\left(\left(\overline{x}-\overline{y}\right)-t_{\alpha/2}(S.E),\left(\overline{x}-\overline{y}\right)+t_{\alpha/2}(S.E)\right)=(-3.663,5.663)$

Exercise:

1. Samples of two types of electric light bulbs were tested for length of life and following data were obtained

Sample No.	Type-I	Type-II
	$n_1=8$	$n_2=7$
Sample Means	$\overline{x}_1=1,234\ hrs.$	$\overline{x}_2=1,2036\ hrs.$
Sample S.D.'s	$s_1=36\ hrs.$	$s_2=40\ hrs.$

Is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life? (Ans: 9.39)

2. Two horses A and B were tested according to the time (in seconds) to run a particular track with the following results .

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	-

Test whether the two horses have the same running capacity.

3. The mean life of a sample of 10 electric bulbs (or motors) was found to be 1456 hours with S.D of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 hours with S.D of 398 hours. Is there a significant difference between the means of two batches?

4. Below are given the gain in weights (in kgs.) of cow fed on two diets A and B.

Gain in weight	
Diet A	25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25
Diet B	44, ,34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Test, if the two diets differ significantly as regards their effect on increase in weight. (Ans:-0.609)

Paired t-test

Paired observations arise in many practical situations where each homogeneous experimental unit receives both population conditions. As a result, each experiment unit has a pair of observations, one for each population.

Consider the case when

(i) the sample sizes are equal $n_1 = n_2 = n$

(ii) The two samples are not independent but the sample observations are paired together i.e., the pair of observations $(x_i, y_i)\ (i = 1, 2, \ldots, n)$ correspond to same sample unit.

Suppose a business concern is interested to know whether a particular media of promoting sales of a product is really effective or not. In this case we have to test whether the average sales before and after the sales promotion are equal.

If $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ be the pairs of sales data before and after the sales promotion of a business concern. We apply t – test to examine the significant difference between the mean in two situations.

Let $d_i = x_i - y_i$ for i =1, 2, 3, …., n
Let the Null Hypothesis be $H_0 : \mu_d = 0$, there is no significant difference between the means in two situations.
Then the Alternative Hypothesis is $H_1 : \mu_d > 0$
Assuming that H$_0$ is true , the test statistic for n paired observations (which are dependent) by taking the differences $d_1, d_2, \ldots\ldots, d_n$ of the paired data.

$$t = \frac{\bar{d} - \mu}{S/\sqrt{n}} = \frac{\bar{d}}{S/\sqrt{n}} \qquad (\because \mu = 0)$$

Where $\bar{d} = \frac{1}{n}\sum d_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$

are the mean and variance of the differences $d_1, d_2, \ldots\ldots, d_n$ respectively and μ is the mean of the population of differences .
The above statistic follows student's t – distribution with (n-1) degrees of freedom.

Example 1: The Blood Pressure of 5 women before and after intake of a certain drug are given below:

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

Test whether there is significant change in Blood Pressure at 1% level of significance.

Solution: Let the Null Hypothesis be $H_0 : \mu_d = 0$, i.e., there is no significant difference in blood pressure before and after intake of a drug. Then Alternative Hypothesis is $H_1 : \mu_d > 0$

Assuming that H₀ is true , the test statistic is $t = \frac{\bar{d}}{S/\sqrt{n}}$ where $\bar{d} = \frac{\sum d}{n}, d = y - x$

Calculation For $\bar{d}$ and S

Women	**B.P before intake of drug(x)**	**B.P after intake of drug(x)**	**d=y-x**	**d²**
1	110	120	10	100
2	120	118	-2	4
3	123	125	2	4
4	132	136	4	16
5	125	121	-4	16
Total			$\sum d$ =10	$\sum d^2$ = 140

$$\therefore \bar{d} = \frac{\sum d}{n} = \frac{10}{5} = 2 \text{ and } S^2 = \frac{\sum (d-\bar{d})^2}{n-1} = \frac{\sum d^2 - (\bar{d})^2 \times n}{n-1} = 30, \quad \therefore S = \sqrt{30}$$

$$\therefore t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{2}{\sqrt{30}/\sqrt{5}} = \frac{2}{\sqrt{6}} = 0.82$$

Degrees of freedom = n-1 = 5-1 = 4

Thus t = 0.82 < 4.6 at 1% level with 4 d.f.

Since the calculated value of t < the tabulated value with 4 d.f. at 1 % level , we accept H₀ at 1% level and conclude that there is no significance change in Blood Pressure after intake of a certain drug.

Exercise:

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

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

(Ans: 1.36, Accept H₀)

2. In a certain experiment to compare two types of animals foods A and B, the following results of increase in weights were observed in animals:

Animal Number		1	2	3	4	5	6	7	8	Total
Increase weight in lb	Food A	49	53	51	52	47	50	52	53	407
	Food B	52	55	52	53	50	54	54	53	423

Examine the case when the same sets of eight animals were used in both the foods.

3. The scores of soldiers before and after training are given as follows

Soldiers Before	67	24	57	55	63	54	56	68	33	43
Soldiers After	70	38	58	58	56	67	68	75	42	38

Do the data indicate that the soldiers have been benefitted by the training. Ans: 2.16, Reject H_0

Applications of t-test:

The t-distribution has a wide number of applications in statistics, some of which are enumerated below.

(i) To test if the sample mean $(\bar{x})$ differs significantly from the hypothetical value μ of the population mean

(ii) To test the significance of the difference between two sample means.

(iii) To test the significance of an observed sample correlation coefficient and sample regression coefficient.

(iv) To test the significance of observed partial correlation coefficient.

Snedecor's F-test for variance:

When testing the significance of the difference of the means of two samples, we assumed that two samples came from the same population or from populations
With equal variances. If the variances of the population are not equal , a significant difference in the means may arise . Hence , before we apply the t – test for the significance of the difference of two means, we have to test for the equality of population variances using F- test of significance.

If s_1^2 and s_2^2 are the variances of two samples of sizes n_1 and n_2 respectively then the population variances are given by

$n_1 s_1^2 = (n_1 - 1)S_1^2$ and $n_2 s_2^2 = (n_2 - 1)S_2^2$

The quantities $v_1 = (n_1 - 1)$ and $v_2 = (n_2 - 1)$ are called degrees of freedom of these estimates .We want to test if these estimates S_1^2 and S_2^2 are significantly different or if the samples may be regarded as drawn from the same population or from two populations with same variance σ^2 .

Test for Equality of Two population variances:

Let two independent random samples of sizes n_1 and n_2 be drawn from two normal populations. To test the hypothesis that the two population variances σ_1^2 and σ_2^2 are equal.

Let the Null hypothesis be $H_0 : \sigma_1^2 = \sigma_2^2$

Then the Alternative hypothesis is $H_1 : \sigma_1^2 \neq \sigma_2^2$

The estimates of σ_1^2 and σ_2^2 are given by

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{\sum (x_i - \bar{x})^2}{n_1 - 1} \text{ and } S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{\sum (y_i - \bar{y})^2}{n_2 - 1},$$

where s_1^2 and s_2^2 are the variances of the two samples.

Assuming that H₀ is true , the test statistic $F = \frac{S_1^2}{S_2^2}$ or $\frac{S_2^2}{S_1^2}$ according as S_1^2 > S_2^2 or S_2^2 > S_1^2

Follows F- distribution with (n_1-1 , n_2-1) degrees of freedom.
Conclusion: If the calculated value of F > the tabulated value of F at 5% level ,we reject the Null hypothesis H₀ and conclude that the variances σ_1^2 and σ_2^2 are not equal.

Otherwise, we accept the Null hypothesis H₀ and conclude that σ_1^2 and σ_2^2 are equal.

Example1: In one sample of 8 observations from a normal population , the sum of the squares of deviations of the sample values from the sample mean is 84.4 and in another sample of 10 observation it was 102.6 .Test at 5% level whether the populations have the same variance.

Solution: Let σ_1^2 and σ_2^2 be the variances of the two normal population from which the samples are drawn .

Let the Null hypothesis be $H_0 : \sigma_1^2 = \sigma_2^2$

Then the Alternative hypothesis is $H_1 = \sigma_1^2 \neq \sigma_2^2$

Here $n_1 = 8$, $n_2 = 10$

Also $\sum (x_i - \bar{x})^2$ = 84.4 , $\sum (y_i - \bar{y})^2$ = 102.6, If S_1^2 and S_2^2 be the estimates of σ_1^2 and σ_2^2

$$S_1^2 = \frac{\sum (x_i - \bar{x})^2}{n_1 - 1} = \frac{84.4}{7} = 12.057 \quad \text{and} \quad S_2^2 = \frac{\sum (y_i - \bar{y})^2}{n_2 - 1} = \frac{102.6}{9} = 11.4$$

Let H₀ be true . Since S_1^2 > S_2^2 , the test statistic is

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

i.e., calculated F = 1.057.
Degrees of freedom are given by $v_1 = (n_1 - 1) = 8-1 = 7$ and $v_2 = (n_2 - 1) = 10 - 1 = 9$.
Tabulated value of F at 5% level for (7,9) degrees of freedom is 3.29
i.e., $F_{0.05}(7,9) = 3.29$
Since calculated F < tabulated F , we accept the Null hypothesis H₀ and conclude that the population have the same variance.

Exercise:

1. Two random samples gave the following results:

Sample	Size	Sample mean	Sum of squares of deviations from the mean
1	10	15	90
2	12	14	108

Test whether the samples come from the same normal population at 5% level of significance.

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

3. The measurements of the output of two units have given the following results. Assuming that both samples have been obtained from the normal populations at 10% significance level, test whether the two populations have the same variance.

Unit-A	14.1	10.1	14.7	13.7	14.0
Unit-B	14.0	14.5	13.7	12.7	14.1

Chi-square (χ^2) test:

If O$_i$ (i = 1,2,...,n) is a set of observed frequencies and E$_i$(i = 1,2,....,n) is the correspondimg set of expected frequencies then χ^2 is defined as $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ with (n-1) degrees of freedom. χ^2 test is used to test whether differences between observed and expected frequencies are significant.

Note: If the given data is given in a series of 'n' numbers then degrees of freedom = n-1.

Chi- square distribution is an important continuous probability distribution and it is used in both large and small tests. In chi-square tests , χ^2 - distribution is mainly used

(i) To test the goodness of fit .
(ii) To test the independence of attributes.
(iii) To test if the population has a specified value of the variance σ^2.

Chi-square test for goodness of fit:

Let the Null hypothesis H$_0$ be that there is no significant difference between the observed values and the corresponding expected values.

Then the Alternative Hypothesis H$_1$ is that the above difference is significant.

Let $O_1, O_2, \ldots\ldots, O_n$ be a set of observed frequencies and $E_1, E_2, \ldots\ldots, E_n$ the corresponding set of expected frequencies .then the test statistic χ^2 is given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \ldots.. + \frac{(O_n - E_n)^2}{E_n}$$

Assuming that H$_0$ is true , the test statistic χ^2 follows Chi-square distribution with (n -1)d.f.,

Conclusion: If the calculated value of χ^2 > tabulated value of χ^2 at α level , the Null hypothesis H$_0$ is rejected . Otherwise , H$_0$ is accepted.

Example1: A die is thrown 264 times with the following results .Show that the die is biased.

No. appeared on the die	1	2	3	4	5	6
Frequency	40	32	28	58	54	52

Solution: Null Hypothesis H_0: The die is unbiased

The expected frequency of each of the numbers 1,2,3,4,5,6 is $\frac{264}{6} = 44$

Calculations for χ^2

Observed frequency (O_i)	**Expected frequency (E_i)**	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
40	44	16	0.3636
32	44	144	3.2727
28	44	256	5.8181
58	44	196	4.4545
54	44	100	2.2727
52	44	64	1.4545
264	**264**		**17.6362**

$\therefore \chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 17.6362$. The number of degrees of freedom = n – 1 =5

The tabulated value of χ^2 for 5 d.f at 5% level = 11.07

Since calculated χ^2 > tabulated χ^2 ,we reject the null hypothesis H_0.

i.e., we reject the hypothesis that the die is unbiased . Hence the die is biased.

Exercise:

1. The demand for a particular spare part in a factory was found to vary from day-to-day. In a sample study the following information was obtained:

Days	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
No. of parts demanded	1124	1125	1110	1120	1126	1115

Test the hypothesis that the number of parts demanded does not depend on the day of the week.(Given the values of chi-square significance at 5,6,7 d.f. are respectively. 11.07, 12.59, 14.07 at the 5% level of significance . (Ans: 0.179)

2. A sample analysis of examination results of 200 MBA's was made. It was found that 46 students had failed , 68 secured a third division, 62 secured a second division and the rest were placed in first division. Are these figures commensurate with the general examination result which is in the ratio of 4 : 3 : 2 : 1 for various categories respectively? (Ans: 28.417)

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

Digits		0	1	2	3	4	5	6	7	8	9
Freequency		1026	1107	997	966	1075	933	1107	972	964	853

Test whether the digits may be taken to occur equally frequently in the directory.

(Ans: 58.542)

Degrees of freedom: If expected frequencies E_i are obtained by fitting of data using

1. Binomial Distribution d.o.f = n-1 (one statistic, p to be found)
2. Poisson Distribution d.o.f = n-2 (two statistics, N and λ to be found)

Exercise:

1. Five unbiased dice are thrown 96 times and the no:of times 4 or 5 or 6 was obtained as below

No. of dice showing 4 or 5 or 6	5	4	3	2	1	0
Frequency	8	18	35	24	10	1

Fit a suitable distribution and test for goodness of fit.

2. the following is the distribution of the hourly no.of trucks arriving at a company 's ware house

Trucks arriving per hour	0	1	2	3	4	5	6	7	8
Frequency	52	151	130	102	45	12	5	1	2

Fit a poisson distribution and test for goodness of fit.

3. A survey of 800 families with four children each revealed the following distribution:

No. of boys	0	1	2	3	4
No.of girls	4	3	2	1	0
No. of families	32	178	290	236	64

Is the result consistent with the hypothesis that male and female births are equally probable ?

(Ans: 19.63)

Chi-square test for independence of attributes:

Literally, an attribute means a quality or characteristic .Examples of attributes are drinking ,smoking ,blindness , honesty, beauty etc,

Let the observations be classified according to two attributes and the frequencies O_i in the different categories be shown in a two –way table, called contingency table .We have to test on the basis of cell frequencies whether the two attributes are independent or not. We take the Null –Hypothesis H_0 that there is no association between the attributes i.e., we assume that the two attributes are independent. The expected frequencies (E_i) of any cell = $\dfrac{Rowtotal \times Columntotal}{Grandtotal}$

The test statistic $\chi^2 = \sum_{i=1}^{n} \dfrac{(O_i - E_i)^2}{E_i}$ approximately follows Chi-aquare distribution with d.f. = (No. of rows - 1) x (No. of columns - 1)

If the calculated value of χ^2 is less than the table value at a specified level of significance then the attributes are independent and do not bear any association. On the other hand , if the calculated value of χ^2 is greater than the table value at a specified level of significance , we say that the results of the experiment do not support the hypothesis, in other words the attributes are associated.

Let us consider two attributes A and B is divided into two classes and B is divided into two Classes. The various cell frequencies can be expressed in the following table known as 2 x 2 contingency table.

a	b	**a+b**
c	d	**c+d**
a+c	**b+d**	**N = a+b+c+d**

The expected frequencies are given by

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

N = a+b+c+d with d.f = (2-1)(2-1) = 1. We use this formula when the expected frequencies are in fractions .

Example1: The following table gives the classification of 100 workers according to gender and nature of work .Test whether the nature of work is independent of the gender of the worker.

	Stable	**Unstable**	**Total**
Males	40	20	60
Females	10	30	40
Total	30	50	100

Solution: Null Hypothesis H_0: The nature of work is independent of the gender of the workers.
Expected frequencies are given in the table:

$\frac{50\times 60}{100} = 30$	$\frac{50\times 60}{100} = 30$	60
$\frac{50\times 40}{100} = 20$	$\frac{50\times 40}{100} = 20$	40
50	50	100

Calculations for χ^2 **:**

Observed frequency (O_i)	**Expected frequency (E_i)**	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
40	30	100	3.333
20	30	100	3.333
10	20	100	5.000
30	20	100	5.000
100	100	$\sum \frac{(O_i - E_i)^2}{E_i}$	16.66

$$\therefore \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 16.66$$

$$\therefore Calculated\, \chi^2 = 16.66$$

Tabulated value of χ^2 (2-1)(2-1) = 1 d.f at 5% level of significance is 3.84.

Since calculated χ^2 > tabulated χ^2, we reject the null hypothesis H_0 i.e., nature of work is not independent of the gender of the workers.

i.e., there is difference in the nature of work on the basis of gender.

Exercise:

1. Out of 8000 graduates in a town 800 are females, out of 1600 graduate employees 120 are females. Use χ^2 to determine if any distinction is made in appointment on the basis of gendre. χ^2 at 5% level for one degree of freedom is 3.84. (Ans: 13.84)

2. Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas. The results are taken in the below table. Examine whether the nature of the area is related to voting preference in this election. (Ans:10.0891)

Area	Votes for		Total
	A	B	
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

3. A random sample of students of XYZ University was selected and asked their opinion about autonomous colleges'. The results are given below. The same number of each gender was included within each class group. Test the hypothesis at 5% level that opinions are independent of the class groupings :

(Ans:13.89)

Class	Numbers		Total
	Favoring 'autonomous colleges'	Opposed to 'autonomous colleges'	
B.A/B.Sc/B.Com. Part I	120	80	200
B.A/B.Sc/B.Com. Part II	130	70	200
B.A/B.Sc/B.Com. Part III	70	30	100
M.A/M.Sc/M.Com.	80	20	100
Total	400	200	600

4. Four methods are under development for making discs of a super conducting material. Fifty discs are made by each method and they are checked for super conductivity when cooled with liquid.

	1st Method	2nd Method	3rd Method	4th Method
Super Conductors	31	42	22	25
Failures	19	8	28	25

Test the significant difference between the proportions of super conductors at 0.05 level

5. From the following data, find whether there is any significant liking in the habit of taking soft drinks among the categories of employees. Use Chi-square distribution test with l.o.s 0.05

Employees

Soft Drinks	**Clerk**	**Teachers**	**Officers**
Pepsi	10	25	65
Thums Up	15	30	65
Fanta	50	60	30

Applications of Chi-Square test:

χ^2 -distribution has a large number of applications in statistics, some of which are enumerated below:

(i) To test if the hypothetical value of the population variance is $\sigma^2 = \sigma_0{}^2$ (say).

(ii) To test the 'goodness of fit'.

(iii) To test the independence of attributes.

(iv) To test the homogeneity of independent estimates of the population variance.

(v) To combine various probabilities obtained from independent experiments to give a single test of significance.

(vi) To test the homogeneity of independent estimates of the population correlation coefficient.