NOTES ON
Probability & Statistics
MCA I Year/ I Semester
(2023-24)

By

DEPAVATH HARINATH
(ASSISTANT PROFESSOR)

DEPARTMENT OF MCA

RAMNATH GULJARILAL KEDIA COLLEGE OF COMMERCE
3-1-336, OPP NEW CHADERGHAT BRIDGE,
Kachiguda Station Road,
Esamia Bazar, Hyderabad 500 027, Telangana,India.

(PCC105)PROBABILITY AND STATISTICS

Course Objectives:

- Understand the Linear Algebra concepts through vector spaces.
- Basic concepts of probability and concepts of various discrete and continuous probability distributions.
- Learning sampling procedure and various kinds of estimate techniques.
- Learning hypotheses testing and acquiring knowledge of basic statistical Inference and its applications.
- The concept of association between two variable and forecast future values by regression equations.

Course Outcomes :

- Understanding of Linear Algebra will boost the ability to understand and apply various data science algorithms.
- Calculate probabilities by applying probability laws and theoretical results, knowledge of important discrete and continuous distributions, their inter relations with real time applications.
- Understanding the use of sample statistics to estimate unknown parameters.
- Become proficient in learning to interpret outcomes.
- Compute and interpret Correlation Analysis, regression lines and multiple regression analysis with applications.

Suggested Readings:

1. David C Lay, Linear Algebra and its Applications 4e.
2. Richard I Levin, David S Rubin – Statistics for Management, Seventh Edition, PHI – 1997.
3. R D Sharma “Theory and Problems of Linear Algebra”, International Publishing House Pvt. Limited, 2011.
4. A K Sharma, “Linear Algebra”, Discovery Publishing House Ltd., 2019.
5. Gilbert Strang, Linear Algebra and its Applications, 2010.
6. S. C. Gupta and V.K.Kapoor, Fundamentals of Mathematical Statistics Sultan Chand & Sons, New Delhi.

Course Material

I MCA

PROBABILITY AND STATISTICS

UNIT I

Probability & Random Variable

Probability Spaces:

Mathematical definition of probability:

If there are n exhaustive, mutually exclusive and equally likely events, probability of the happening of A is defined as the ration m/n, m is favourable to A.

$$P(A) = \frac{m}{n}$$

$$= \frac{\text{Number cases favourable to A}}{\text{Exhaustive number of cases in S}}$$

Statistical definition of probability:

Let a random experiment be repeated n times and let an event A occur n_A out of n trials. The ratio $\frac{n_A}{n}$ called the relative frequency of the event A. As n increases, $\frac{n_A}{n}$ shows a tendency to stabilize and to approach a constant value. This value, denoted by P(A) is called the probability of the event A.

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

Axiomatic definition of Probability:

Let S be the sample space and A be an event associated with a random experiment. Then the probability of the event A, denoted by P(A), is defined as a real number satisfying the following axioms,

(i) $0 \le P(A) \le 1$

(ii) P(S)=1

(iii) If A and B are mutually exclusive events, $P(AUB) = P(A) + P(B)$ and

(iv) If $A_1, A_2, ..., A_n, ...$ are a set of mutually exclusive events,

$$P(A_1 \cup A_2 \cup .... \cup A_n ....) = P(A_1) + P(A_2) + ... + P(A_n) ....$$

Mutually exclusive events:.

When the occurrence of one event procludes the occurrence of all other events, then such a set of events is said to be mutually exclusive.

Example On tossing a coin , either head or tail can occur but not both. i.e occurrence of head excludes the occurrence of tail. The events of occurrence of head and tail are mutually exclusive.

Equally likely events

Two events are said to be equally likely events if each one of them has an equal chance of occurrence.

In tossing an unbiased coin the occurrence of head or tail are equally likely.

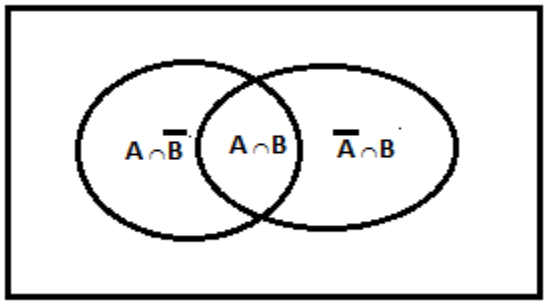
Addition law of probability:

If A and B are any two events, and are not disjoint, then

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

Proof:

From the Venn diagram, the events A and $\bar{A} \cap B$ are disjoint.



Therefore $A \cup B = A \cup (\bar{A} \cap B)$

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

Adding and subtracting $P(A \cap B)$,

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

Conditional Probability

Ans: The conditional probability of an event B, assuming that the event A has happened.

$$P(B/A) = \frac{P(A \cap B)}{P(A)}, provided P(A) \neq 0$$

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

Multiplication law of probability

If the events A and B are independent,

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

Theorem of total probability

If $B_1, B_2, ..., B_n$ be a set of exhaustive and mutually exclusive events, and A is another event associated with B_i, then

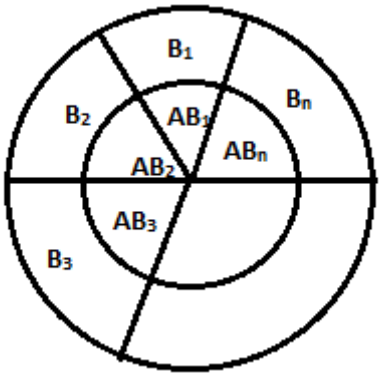
$$P(A) = \sum_{i=1}^{n} P(B_i)P(A/B_i)$$

Proof:

The inner circle represents the event A.

A can occur along with $B_1, B_2, ..., B_n$ that are exhaustive and mutually exclusive.

Therefore $AB_1, AB_2, ..., AB_n$ are also mutually exclusive.



$$A = AB_1 + AB_2 + ... + AB_n$$

$$P(A) = P(\sum AB_i)$$

$$P(A) = \sum P(AB_i)$$

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

Bayes theorem

If $B_1, B_2, ..., B_n$ be a set of exhaustive and mutually exclusive events, and A is another event associated with B_i, then

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

Proof:

$$P(B_i \cap A) = P(B_i) \times P(A/B_i)$$

$$P(B_i \cap A) = P(A) \times P(B_i/A)$$

$$\therefore P(A/B_i) = \frac{P(B_i) \times P(A/B_i)}{P(A)}$$

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

Random Variable:

A real – valued function defined on the outcome of a probability experiment is called a random variable.

Example : Suppose that a coin is tossed twice so that the sample space is $S = \{HH, HT, TH, TT\}$. Let X represent the number of heads that can come up. With each sample point we can associate a number for X as shown in Table . Thus, for example, in the case of HH (i.e., 2 heads), $X = 2$ while for TH (1 head), $X = 1$. It follows that X is a random variable.

Table

Sample Point	*HH*	*HT*	*TH*	*TT*
X	2	1	1	0

It should be noted that many other random variables could also be defined on this sample space, for example, the square of the number of heads or the number of heads minus the number of tails.

Discrete Random Variable :

A random variable whose set of possible values is either finite or countably infinite is called discrete random variable.

Example: Number of transmitted bits received in error.

Cumulative Distribution Function :

The cumulative distribution F(x) of a discrete random variable X with probability distribution f(x) is given by

$F(x) = P(X \le x) = \sum_{t \le x} f(t)$ for $-\infty < x < \infty$

Mean or Expectation of a discrete Random variable X:

Let X be a discrete random variable assuming values $x_1, x_2, \ldots, x_n$ with corresponding probabilities $P_1, P_2, \ldots, P_n$. Then

$E(X) = \sum_i x_i p(x_i)$ is called the expected value of X.

E(X) is also commonly called the mean or the expectation of X. A useful identity states

that for a function g,

$E[g(x)] = \sum_{x_i} g(x_i)\, p(x_i)$

Continuous Random Variable:

A random variable X is said to be continuous if it takes all possible values between certain limits say from real number ‘a’ to real number ‘b’.

Example: The length of time during which a vacuum tube installed in a circuit functions is a continuous random variable, number of scratches on a surface, proposition of defective parts among 1000 tested, number of transmitted in error.

Cumulative distribution function of a continuous random variable:

The cumulative distribution function of a continuous random variable X is

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\, dt \text{ for } -\infty < x < \infty$$

Mean or Expectation of a Continuous Random variable X:

Suppose X is a continuous random variable with probability density function f(x). The mean or expected value of X, denoted as μ or E(X) is

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

A useful identity is that for any function g,

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$$

Variance of X:

The variance of X, denoted as V(X) or σ^2, is

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

Moment generating function of a random variable X about the origin:

Moment generating function of a random variable X about the origin is defined as

$$M_X(t) = E[e^{tX}] = \sum_x e^{tx} p(x) \text{ , if X is discrete}$$
$$\int_{-\infty}^{\infty} e^{tx}\, f_X(x)\, dx \text{ , if X is continuous.}$$

Tchebyshev's Inequality:

Let X be a random variable with mean $E(X) = \mu$ and variance $var(X) = \sigma^2$. Then the Tchebyshev's inequality states that

$$P\left(|X-\mu| \ge t\right) \le \frac{\sigma^2}{t^2}$$

For any t>0.

Other equivalent form can be written for this inequality is

$$P\left(|X-\mu| < t\right) \le 1 - \frac{\sigma^2}{t^2}$$

$$P\left(|X-\mu| \ge n\sigma\right) \le \frac{1}{n^2}$$

Problems:

1. Find the chance of throwing (a) four (b) an even number with an ordinary six faced die.

   P (throwing four) = $\frac{1}{6}$

   P (getting an even number) = $\frac{3}{6} = \frac{1}{2}$.

2. A bag contains 8 white balls and 6 red balls. Find the probability of drawing two balls of the same colour.

   Two balls out of 14 balls can be drawn in $14C_2$ ways.

   Two white balls out of 8 can be drawn in $8C_2$ ways.

   P (drawing two white balls) = $\frac{8C_2}{14C_2} = \frac{28}{91}$

   Two red balls out of 6 can be drawn in $6C_2$ ways.

   P (drawing two red balls) = $\frac{6C_2}{14C_2} = \frac{15}{91}$

   Probability of drawing 2 balls of same colour (either both white or both red)

   = $\frac{28}{91} + \frac{15}{91} = \frac{43}{91}$

3. Find the probability of drawing an ace or a spade or both from a deck of cards?

   Probability of drawing an ace event (A) = $\frac{4}{52}$.

   Probability of drawing a spade event (B) = $\frac{13}{52}$

   Probability of drawing an ace of spade ($A \cap B$) = $\frac{1}{52}$.

   The events drawing an ace or a spade are not mutually exclusive , therefore

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

   = $\frac{4}{52} + \frac{13}{52} - \frac{1}{52}$

   = $\frac{4}{13}$.

4. What is the chance that a leap year selected at random will contain 53 Sundays?

   A leap year contains 52 full weeks and extra two days (total of 366 days).

   Possible two days combinations are

   Monday & Tuesday

   Tuesday & Wednesday

   Wednesday & Thursday

   Thursday & Friday

   Friday & Saturday

Saturday & Sunday

Sunday & Monday

There 7 possible combinations. We have Sunday in two combinations

The required probability $= \frac{2}{7}$

5. When A and B are 2 mutually exclusive events such that P(A) = ½ and P(B) = 1/3 , find P(AUB) and P(A$\cap B$).

P(AUB) = P(A) +P(B) = ½ +1/3 = 5/6

P(A$\cap B$) = 0

6. A fair coin is tossed 5 times what is the probability of having at least one head?

n(S) = 2^5 =32

n(A) = at least one head = 31

P(A) = 31/32

7. Given that P(A) =0.31, P(B) = 0.47 , A and B are mutually exclusive. Then find P(A$\cap \bar{B}$).

A and B are mutually exclusive, therefore P(A$\cap B$) = 0

W.K.T P(A$\cap \bar{B}$) + P(A$\cap B$) $= P(A)$

P(A$\cap \bar{B}$) $= 0.31$.

8. If P(A) =0.35 P(B) = 0.13 and P(A$\cap B$) $= 0.14$ find $P(\bar{A} \cup \bar{B})$.

$P(\bar{A} \cup \bar{B})$ = P$\overline{A \cap B}$ = 1- P(A$\cap B$) = 0.86.

9. Given P(A) =1/3 , P(B) =1/4 , P(A$\cap B$) $= 1/6$ find the following probability $P(\bar{A})$, $P(\bar{A} \cap \bar{B})$.

$P(\bar{A})$= 1- P(A) = 2/3

$P(\bar{A} \cap \bar{B})$.= P($\overline{AUB}$). = 1- P(AUB) = 7/12.

10. What is the probability of obtaining 2 heads in two throws of a single coin?

n(S) = 4 n(A) =1 : P(A) = ¼.

11. If P(A)=P(B)=P(AB), show that $P(A\bar{B}+\bar{A}B)=0$

Soln:

By addition theorem,

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

We can write,

$P(A\cup B)=P(AB)+P(\bar{A}B)+P(A\bar{B})$ -----(2)

Using the given condition in (1),

$P(A\cup B)=P(AB)$ ----------(3)

From (2) and (3), $P(\bar{A}B)+P(A\bar{B})=0$

12. A box contains tags marked 1,2,…,n. two tags are chosen at random without replacement. Find the probability that the numbers on the tags will be consecutive integers.

Soln:

No. of ways of choosing any one pair from the (n-1)pairs = $(n-1)C_1=n-1$

Total no. of ways of choosing 2 tags from the n tags = nC_2

Therefore the required probability = $\frac{n-1}{n(n-1)/2}=2/n$

13. Among the workers in a factory only 30% receive a bonus. Among those receiving the bonus only 20% are skilled. What is the probability of a randomly selected worker who is skilled and receiving bonus.

Soln:

P(A)=0.3

P(B/A)=0.2

$P(A\cap B)=P(A)P(B/A)$

= 0.6

14. Prove that the events A and B are independent, then $\bar{A}$ and $\bar{B}$ are also independent.

Proof: $P(\bar{A}\cap\bar{B})=P(\overline{A\cup B})$

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

Using addition and multiplication theorems,

$P(\bar{A}\cap\bar{B})=P(\bar{A})\times P(\bar{B})$

15. A and B alternately throw a pair of dice. A wins if he throws 6 before B throws 7 and B wins if he throws 7 before A throws 6. If A begins, show that his chance of winning is 30/61.

Soln: A- Event of A throwing 6.

B – Event of B throwing 7.

$P(A)=\frac{5}{36}$ $P(B)=\frac{1}{6}$

$P(Awins)=P(Aor\bar{A}\bar{B}Aor\bar{A}\bar{B}\bar{A}\bar{B}Aor...)$

$=P(A)+P(\bar{A}\bar{B}A)+P(\bar{A}\bar{B}\bar{A}\bar{B}A)+...$

= 30/61

16. In a coin tossing experiment, if the coin shows head, 1 dice is thrown and the result is recorded. But if the coin shows tail, 2 dice are thrown and their sum is recorded. What is the probability that the recorded number will be 2?

Soln:

When a single die is thrown, P(2)=1/6.

When 2 dice are thrown, the sum will be 2, only if each dies shows 1.

Therefore P(getting 2 as sum with 2 dice) =1/6 X 1/6=1/36

by theorem of total probability,

$$P(2) = P(H) \times P(2/H) + P(T) \times P(2/T)$$

= 7/72.

17. If atleast 1 child in a family with 2 children is a boy, what is the probability that both children are boys?

Soln:

P=probability that a child is a boy = ½

q=1/2

P(atleast one boy) = P(exactly 1 boy)+P(exactly 2 boys)

= ¾

P(both are boys/atleast one is a boy)= $\frac{\frac{1}{4}}{\frac{3}{4}} = \frac{3}{4}$

18. In a shooting test, the probability of hitting the target is ½ for A, 2/3 for B and ¾ for C. If all of them fire at the target, find the probability that none of them hits the target.

Soln:

Let A,B, and C are the event of hitting the target.

P(A)=1/2 ; P(B)=2/3 ; P(C)= ¾

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

=1/24

19. If $\overline{A}$ is the complementary event of A, prove that $P(\overline{A}) = 1 - P(A) \leq 1$

Proof: If A and are mutually exclusive events, such that $A \cup \overline{A} = S$

$$P(A \cup \overline{A}) = P(S)$$

$$P(A) + P(\overline{A}) = 1$$

$$P(\overline{A}) = 1 - P(A)$$

Since $P(A) \geq 0$, it follows that $P(\overline{A}) \leq 1$

20. Two fair dice are thrown independently. Three events A, B and C are defined as follows.

(i) Odd face with the first die

(ii) Odd face with second die

(iii) Sum of the numbers in 2 dice is odd. Are the events A, B and C mutually independent?

Soln:

P(A)= ½; P(B)= ½; P(C)=1/2

$$P(A \cap B) = P(B \cap C) = P(A \cap C) = 1/4$$

$$P(A \cap B \cap C) = 0$$

Since C cannot happen when A and B occur. Therefore

$P(A \cap B \cap C) \neq P(A)P(B)P(C)$

Therefore the events are pairwise independent, but not mutually independent.

21. Two defective tubes get mixed up with 2 good ones. The tubes are tested, one by one, until both defectives are found. What is the probability that the last defective tube is obtained on (i) the second test (ii) the third test and (iii) the fourth test.

Soln:

Let D represent defective and N represent non-defective tube.

(i) P(Second D in the II test)=P(D in the I test and D in the II test)

$= P(D_1 \cap D_2)$

$= P(D_1) \times P(D_2) = 1/6$

(ii) P(Second D in the III test) $= P(D_1 \cap N_2 \cap D_3 or N_1 \cap D_2 \cap D_3) = 1/3$

(iii) P(Second D in the IV test)

$= P(D_1 \cap N_2 \cap N_3 \cap D_4 or N_1 \cap D_2 \cap N_3 \cap D_4 or N_1 \cap N_2 \cap D_3 \cap D_4) = 1/2$ =

22. If the events A and B are independent then prove that

(i) $\overline{A}$ and $\overline{B}$ are independent.

(ii) $\overline{A}$ and B are independent.

(iii) A and $\overline{B}$ are independent.

Proof: (i) by Demorgan's law

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

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

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

$= 1 - [P(A) + P(B) - P(A \cap B)]$

$= 1 - [P(A) + P(B) - P(A)(B)]$

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

Therefore $\overline{A}$ and $\overline{B}$ are independent.

(ii) the events $A \cap B$ and $\overline{A} \cap B$ are mutually exclusive.

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

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

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

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

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

Therefore $\overline{A}$ and B are independent.

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

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

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

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

Therefore A and $\overline{B}$ are independent.

23. Show that 2^n-(n+1) equations are needed to establish the mutual independence of n events.

Soln: n events are mutually independent, if they are totally independent when considered in set of 2,3,…,n events.

Sets of r events can be chosen from the n events in $^{n}C_{r}$ ways.

To establish total independence of r events.

Say $A_1, A_2,\ldots,A_r$ chosen in any one of the $^{n}C_{r}$ ways.

$$P(A_1, A_2, ..., A_r) = P(A_1) \times P(A_2) \times ... \times P(A_r)$$

Therefore to establish total independence of all the $^{n}C_{r}$ sets, each of r events, we need $^{n}C_{r}$ equations.

Therefore the no.of equations required to establish mutual independence $\sum_{r=2}^{n} {}^{n}C_{r}$

$$= nC_0 + nC_1 + nC_2 + ... + nC_n - (1+n)$$
$$= (1+1)^n - (1+n)$$
$$= (2)^n - (1+n)$$

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

Soln:

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

Let D be the event of the item being defective.

P(A)=1/2 , P(B)=P(C)=1/4

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

=2/100

P(D/C)=4/100

By theorem total of probability,

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

= 1/40

25. A bag contains 5 red and 3 green balls and a second bag 4 red and 5 green balls. One of the bags is selected at random and a draw of 2 balls is made from it. What is the probability that one of them red and the other is green.

Soln:

$P(A_1)=P(A_2)=1/2$

B denote the event of selecting one red and one green ball.

$P(B/A_1)=15/28$

$P(B/A_2)=5/9$

The required probability = $P(A_1)\ P(B/A_1) + P(A_2)\ P(B/A_2)$

= 275/504

26. An urn contains 10 white and 3 black balls. Another urn contains 3 white and 5 black balls. Two balls are drawn at random from the first urn and placed in the second urn and then 1 ball is taken at random from the latter. What is the probability that it is a white ball?

Soln:

The two balls transferred may be both white or both black or 1 white and 1 black.

Let B_1 be the event of drawing 2 white balls from the first urn and B_2 be the event of drawing 2 black balls from it and B_3 be the event of drawing 1 white and 1black ball from it.

Let A be the event of drawing a white ball from the second urn after transfer.
$P(B_1)$= 15/26, $P(B_2)$=1/26, $P(B_3)$=10/26,
$P(A/B_1)$=P(drawing a white ball/2 white balls have been transferred)''
= 5/10.
Similarly, $P(A/B_2)$=3/10 and $P(A/B_3)$=4/10

Therefore $P(A) = P(B_1) \times P(A/B_1) + P(B_2) \times P(A/B_2) + P(B_3) \times P(A/B_3)$
= 59/130

27. A bag contains 5 balls and it is not known how many of them are white. Two balls are drawn at random from the bag and they are noted to be white. What is the chance that all the balls in the bag are white?

Soln:

since 2 white balls have been drawn out, the bag must have contained 2, 3, 4 or 5 white balls.

Let B_1 be the event of the bag containing 2 white balls, B_2 be the event of the bag containing 3 white balls, B_3 be the event of the bag containing 4 white balls and B_4 be the event of the bag containing 5 white balls.

Let A be the event of drawing 2 white balls.
$P(A/B_1)$=1/10; $P(A/B_2)$=3/10; $P(A/B_3)$=3/5; $P(A/B_4)$=1
$P(B_1)$= $P(B_2)$= $P(B_3)$= $P(B_4)$=1/4
By Bayes theorem,

$$P(B_4/A) = \frac{P(B_4)P(A/B_4)}{\sum_{i=1}^{4} P(B_i)P(A/B_i)}, i = 1,2,3,4$$

=1/2.

28. In a bolt factory, machines A,B and C produce 25, 35 and 40% of the total output respectively. Of their outputs 5, 4 and 2% respectively are defective bolts. If a bolt is chosen at random from the combined output, what is the probability that it is defective? If a bold chosen at random is found to be defective, what is the probability that it was produced by B?

Soln:

$P(E_1) = 0.25$; $P(E_2) = 0.35$; $P(E_3) = 0.40$
Let X be the event of drawing defective bolt.
$P(X/E_1) = 0.05$
$P(X/E_2) = 0.04$
$P(X/E_3) = 0.02$
By Baye's theorem

$$P(E_2/X) = \frac{P(E_2)P(X/E_2)}{P(E_1)P(X/E_1) + P(E_2)P(X/E_2) + P(E_3)P(X/E_3)}$$

=0.406.

29. The contents of three urns 1, 2, and 3 are as follows:

Urns Balls	White	Black	Red
I	1	2	3
II	2	3	1
III	3	1	2

An urn is chosen at random and from it two balls are drawn at random. The two balls are one red and one white. What is the probability that they come from the second urn.

Soln:

$$P(B_1) = P(B_2) = P(B_3) = \frac{1}{3}$$

$$P(A/B_1) = \frac{2}{15}$$

$$P(A/B_2) = \frac{2}{5}$$

$$P(A/B_3) = \frac{1}{5}$$

By Baye's theorem,

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

$$P(B_2/A) = \frac{P(B_2)P(A/B_2)}{\sum_{i=1}^{3} P(B_i)P(A/B_i)}, i = 1,2,3$$

= 2/11

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

Soln:

E be the event of chip is actually good and D be the event of tester says it is good.

$$P(\overline{E}) = 0.02$$

$$P(E) = 1 - P(\overline{E}) = 0.98$$

Given that the probability of tester says the chip is good when it is really good is 0.95

$$P(D/E) = 0.95$$

$$P(\overline{D}/E) = 1 - P(D/E) = 0.05$$

$$P(\overline{D}/\overline{E}) = 0.94$$

The probability of actually defective

By Baye's theorem,

$$P(\overline{E}/\overline{D}) = \frac{P(\overline{E}/\overline{D})P(\overline{E})}{P(\overline{E}/\overline{D})P(\overline{E}) + P(\overline{D}/E)P(E)}$$

=0.2773.

31. A certain firm has plant A, B and C producing IC chips. Plant A produces twice the output from B and B produces twice the output from C. The probability of a non-defective product produced by A,B and C are respectively 0.85, 0.75 and 0.95. A customer receives a defective product. Find the probability that it came from plant B.

Soln:

P(A)=1; P(B)=0.5; P(C)=0.25

P(E/A)=0.85 ; P(E/B)=0.75 ; P(E/C)=0.95

$P(\overline{E}/A)=0.15$

$P(\overline{E}/B)=0.25$

$P(\overline{E}/C)=0.05$

The probability that the customer receives a defective product from plant B is

$$P(B/\overline{E})=\frac{P(B)P(\overline{E}/B)}{P(A)P(\overline{E}/A)+P(B)P(\overline{E}/B)+P(C)P(\overline{E}/C)} = 0.4348$$

32. There are 3 true coins and 1 false coin with 'head' on both sides. A coin is chosen at random and tossed 4 times. If 'head' occurs all the 4 times, what is the probability that the false coin has been chosen and used?

Soln:

P(T)=P(the coin is a true coin)=3/4

P(F)=P(the coin is a false coin)=1/4

Let A be the event of getting all heads in 4 tosses.

$$P(A/T)=\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{16}$$

$P(A/F)=1$

By Baye' theorem,

$$P(F/A)=\frac{P(F)\times P(A/F)}{P(F)\times P(A/F)+P(T)\times P(A/T)}$$

= 16/19

33. A coin with is tossed n times. Show that the probability that the number of heads obtained is even is $0.5\left[1+(q-p)^n\right]$.

Soln:

P(even no.of heads are obtained)=P(0 head or 2 head or 4 head or …)

=P(0 head or 2 head or 4 heads or …)

$=nC_0q^np^0+nC_2q^{n-2}p^2+nC_4q^{n-4}p^4+...$ --------(1)

$(q+p)^n=nC_0q^np^0+nC_1q^{n-1}p^1+nC_2q^{n-2}p^2+nC_3q^{n-3}p^3+nC_4q^{n-4}p^4+...$-------(2)

$(q-p)^n=nC_0q^np^0-nC_1q^{n-1}p^1+nC_2q^{n-2}p^2-nC_3q^{n-3}p^3+nC_4q^{n-4}p^4+...$ -------(3)

Adding (2) and (3),

$1+(q-p)^n==2[nC_0q^np^0+nC_2q^{n-2}p^2+nC_4q^{n-4}p^4+...]$ --------(4)

Using (4) in (1),

The required probability = $0.5\left[1+(q-p)^n\right]$

34. Let X be a discrete RV whose cumulative distribution function is

$$F(x)=\begin{cases}0 & \text{for } x<-3\\ 1/6 & \text{for } -3\le x\le 6\\ 1/2 & \text{for } 6\le x\le 10\\ 1 & \text{for } x\ge 10\end{cases}$$

i) Find $P(X\le 4), P(-5<X\le 4)$,

ii) Find the probability distribution of X.

Soln:

a) $P(X\le 4)=F(4)=\frac{1}{6}$

$P(-5<X\le 4)=\frac{1}{6}$

b) The probability distribution of X is

X:	0	-3	6	10
F(X):	0	1/6	1/2	1
P(X):	0	1/6	2/6	1/2

35. The monthly demand for Titan watches is known to have the following probability distribution.

Demand	1	2	3	4	5	6	7	8
Probability	0.08	0.12	0.19	0.24	0.16	0.10	0.07	0.04

Determine the expected demand for watches. Also compute the variance.

Soln:

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

$=1(0.08)+2(0.12)+3(0.19)+4(0.24)+5(0.16)+6(0.10)+7(0.07)+8(0.04)$

$E[X]=4.06$

$E(X^2)=\sum_i x_i^2 p(x_i)$

$=1^2(0.08)+2^2(0.12)+3^2(0.19)+4^2(0.24)+5^2(0.16)+6^2(0.10)+7^2(0.07)+8^2(0.04)$

$=19.7$

$V(X)=E[X^2]-[E(X)]^2$

$=19.7-(4.06)^2=3.2164$

36. If X has the distribution function

$$F(X)=\begin{cases}0 & X<1\\ \frac{1}{3} & \text{for } 1\le X<4\\ \frac{1}{2} & \text{for } 4\le X<6\\ \frac{5}{6} & \text{for } 6\le X<10\\ 1 & \text{for } X\ge 10\end{cases}$$

Find

(i) The probability distribution of X.

(ii) $P(2 < X < 6)$

(iii) Mean of X

(iv) Variance of X

Soln:

(i)

X	0	1	4	6	10
F[x]	0	1/3	½	5/6	1
P(X)	0	1/3	1/6	2/6	1/6

(ii) $p(2 < X < 6) = p[X = 4] \quad = \frac{1}{6}$

(iii) Mean of X = $E[X] = \sum x_i p(x_i) \quad = \frac{28}{6}$

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

$$= 0+1^2(1/3)+4^2(1/6)+ 6^2(2/6)+ 10^2(1/6)$$

$$= 154/6$$

(iv)Variance of X = $Var(X) = E[X^2] - [E[X]]^2$

$$= \frac{190}{6} - \frac{784}{36}$$

$$= \frac{356}{36}$$

$$= \frac{89}{9}$$

37. If $P(X = x) = \begin{cases} Kx, & x = 1,2,3,4,5 \text{ represents a p.m.f} \\ 0, & \text{otherwise} \end{cases}$

(i) Find 'K'

(ii) Find $P(X \text{ being a prime number})$

(iii) Find $P\left\{\frac{1}{2} < X < \frac{5}{2} / X > 1\right\}$

(iv) Find the distribution function.

Soln:

(i) K+2K+3K+4K+5K=1

15K=1

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

(ii) $P(X = x \text{ being a prime number}) = P(X = 2) + P(X = 3) + P(X = 5)$

$$= \frac{2}{15} + \frac{3}{15} + \frac{5}{15} = \frac{10}{15}$$

$$= \frac{2}{3}$$

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

$$=\frac{\frac{2K}{15}}{\frac{2K}{15}+\frac{3K}{15}+\frac{4K}{15}+\frac{5K}{15}}$$

$=1/7$

(iv) The Distribution function $F(x)=P(X\le x)$

$F(x)=0 \quad ; \quad x<1$

$=\frac{1}{15} \quad ; \quad 1\le x<2$

$=\frac{3}{15} \quad ; \quad 2\le x<3$

$=\frac{6}{15} \quad ; \quad 3\le x<4$

$=\frac{10}{15} \quad ; \quad 4<x<5$

$=1 \quad ; \quad 5\le x$

38. a) A fair coin is tossed three times. Let X be the number of tails appearing. Find the probability distribution of X. And also calculate E (X).

b) A continuous random variable X has probability density function given by $f(x)=3x^2, 0\le x\le 1.$ Find K such that $P(X>K)=0.05$

Soln:

a) Let X be an event getting tail, Probability of X is

X	0	1	2	3
P(X)	1/8	3/8	3/8	1/8

$E(X)=\frac{3}{2}$

b) $\int_k^1 f(x)dx=0.05$

$k=(0.95)^{1/3}=0.9830$

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

b) Find the value of (a) C and (b) mean of the following distribution given

$$f(x) = \begin{cases} C\left(x - x^2\right) \text{ for } 0 < x < 1 \\ 0 \quad \text{otherwise} \end{cases}$$

Soln:

a) $p[X < 4] = \int_2^4 f(x)\,dx$

$$k = \frac{2}{27}$$

$$P[X < 4] = \frac{16}{27}$$

b) C = 6

$$\text{Mean} = E[X] = \frac{1}{2}$$

40. A continuous r.v. has the pdf of $f(x) = kx^4$; $-1 < x < 0$. Find the value of k and also

$$P\left(X > -\frac{1}{2} / X < -\frac{1}{4}\right)$$

Soln:

k = 5

$$P\left(X > -\frac{1}{2} / X < -\frac{1}{4}\right) = 0.0303$$

41. A random variable X has density function given by $f(x) = \begin{cases} \frac{1}{k} & \text{for } 0 < x < k \\ 0 & \text{otherwise} \end{cases}$

Find (i) m.g.f. (ii) r^{th} moment (iii) mean (iv) variance.

Soln:

$$M_X(t) = 1 + \frac{kt}{2!} + \ldots \frac{(kt)^r}{(r+1)!} + \ldots$$

$$\text{Coefficient of } t^r = \frac{K^r}{(r+1)!}$$

$$\text{Mean} = \frac{K}{2}$$

$$\text{Variance} = \frac{K^2}{12}$$

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

Soln:

Mean = $A + \mu_1' = 5$

(Variance) $\mu_2 = \mu_2' - \mu_1'^2 = 3$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 0$

$\mu_4 = \mu_4' + 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 26$

43. Find the probability distribution of the total number of heads obtained in four tosses of a balanced coin. Hence the MGP of X, mean of X and variance of X.

Soln:

X:	Number of heads in 4 tosses of a coin				
x:	0	1	2	3	4
p(x):	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

$M_X(t) = E[e^{tX}] = \sum_x e^{tx}p(x)$, if X is discrete

$M_X(t) = \frac{1}{16}[1 + 4e^t + 6e^{2t} + 4e^{3t} + e^{4t}]$

$E[X] = 2$

$\text{Variance}[X] = E(X^2) - [E(X)]^2 = 1$

PROBABILITY DISTRIBUTIONS

Introduction

While constructing probabilistic models for observable phenomena, certain probability distributions arise more frequently than do others. we treat such distributions that play important roles in many engineering applications as special probability distributions.

DISCRETE DISTRIBUTIONS

Bernoulli Trials and Bernoulli Distributions

Let A be an event ((trail) associated with a random experiment such that p(A) remains the same for the repetitions of that random experiment, then the events are called Bernoulli trails.

A random variable X which takes only two values either 1 (success) or 0(failure) with probability p and q respectively. i.e., P(X=1)=p, P(X=0)=q, p+q=1 is called Bernoulli variate and is said to have a Bernoulli distribution.

Definition.

A random variable X is said to follow binomial distribution denoted by B(n,p) if it assumes only non-negative values and its probability mass function is given by

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

=0, otherwise

Where n and p are parameters.

Binomial Frequency Distribution

Suppose that n trails constitute an experiment and if this experiment is repeated N times the frequency function of the binomial distribution is given by

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

Properties of Binomial Frequency Distribution

1. Each trail results in two mutually disjoint outcomes, termed success and failure.

2. The trails must be independent of each other.

3. All trails have same constant probability of success.

4. The number of trails n is finite.

Mean of Binomial Distributions

Mean = $E(X) = \sum_x xp(x)$

$= \sum_{x=0}^{n} x n_{C_x} p^x q^{n-x}$

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

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

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

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

$$= np\sum_{x=1}^{n}(n-1)_{c_{x-1}}\,p^{x-1}q^{n-x}$$

$$= np\sum_{x=1}^{n}(n-1)_{c_{x-1}}\,p^{x-1}q^{(n-1)-(x-1)}$$

$$= np(q+p)^{n-1}$$

Mean=np

Variance of Binomial Distribution

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

The probability mass function of binomial distribution is

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

$$E(X^2) = \sum_{x=0}^{n} x^2 p(x)$$

$$= \sum_{x=0}^{n} x^2 n_{c_x}\,p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x^2\frac{n!}{x!(n-x)!}n_{c_x}\,p^x q^{n-x}$$

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

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

$$= \sum_{x=0}^{n}\frac{n(n-1)(n-2)!}{(x-2)!(n-x)!}\,p^2 p^{x-2}q^{n-x} + E(X)$$

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

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

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

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

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

$$= p^2 \left[n^2 - n - n^2 \right] + np$$

$$= np(1-p)$$

$$= npq$$

Moment Generating Function (M.G.F)

The probability mass function of a binomial distribution is

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

Where n is the number of independent trials and x is the number of success.

By definition of the moment generating function

$$M_x(t) = E(e^{tx})$$

$$= \sum_{x=0}^{n} e^{tx} n_{c_x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} n_{c_x} (pe^t)^x q^{n-x}$$

$$= q^n + n_{c_1} (pe^t)^1 q^{n-1} + n_{c_2} (pe^t)^2 q^{n-2} + ... + (pe^t)^n$$

$$= (q + pe^t)^n$$

Examples

1. The mean and variance of a binomial distribution are 4 and 4/3 respectively. Find P(X≥1) if n=6.

Solution

Mean of binomial distribution = np = 4

Variance of binomial distribution = npq = 4/3

$$\frac{npq}{np} = \frac{\frac{4}{3}}{4}$$

$$q = 1/3$$

Now p = 1-q = 1-1/3 = 2/3

Given n=6

$$P(X = x) = n_{C_x} p^x q^{n-x}$$

$$P(X \geq 1) = 1 - P[X < 1]$$

$$= 1 - P[X = 0]$$

$$= 1 - 6_{c_0} p^0 q^{6-0}$$

$$= 1 - q^6$$

$$= 1 - (\frac{1}{3})^6$$

$$= 1 - \frac{1}{729}$$

$$= \frac{728}{729}$$

2. The mean and variance of binomial distributions are 4 and 3 respectively. Find P(X=0), P(X=1) and P(X≥2).

Solution

Mean of binomial distribution = np = 4

Variance of binomial distribution = npq= 3

$$\frac{npq}{np} = \frac{3}{4}$$

$$q = \frac{3}{4}$$

Now p = 1-q = 1-3/4 = 1/4

Since Mean = np = 4

= n(1/4) = 4

n = 16

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

$$P(X = 0) = n_{c_0}\, p^0 q^n$$

$$= 16_{c_0} (\frac{3}{4})^{16}$$

$$= (\frac{3}{4})^{16} = 0.01$$

$$P(X = 1) = n_{c_1}\, p^1 q^{n-1}$$

$$= 16_{c_1}\, p^1 q^{15}$$

$$= 16(\frac{1}{4})(\frac{3}{4})^{15} = 0.053$$

$$P(X \geq 2) = 1 - P(X < 2)$$

$$= 1 - [P(X = 0) + P(X = 1)]$$

$$= 1 - [0.01 + 0.053] = 1 - 0.063$$

$$= 0.937$$

3. If the mean is 3 and variance is 4 of a random variable X, check whether X follows binomial distribution,

Solution

No. Because for a binomial distribution mean should be greater than the variance.

If mean = np = 3 and variance = npq = 4

npq/np = q = 4/3 = 1.33

1.33 is greater than 1

q>1 (but the probability is less than 1)

Therefore mean should be greater than the variance for a binomial distribution.

3. A binomial variate X satisfies the relation 9P(X=4) = P(X=2) when n=6. Find the parameter p of the binomial distribution.

Solution

The probability function for a binomial distribution is

$$P(X = x) = n_{c_x} p^x q^{n-x}$$

$$P(X = 4) = 6_{c_4} p^4 q^{6-4}$$

$$P(X = 4) = 6_{c_4} p^4 q^2$$

$$P(X = 2) = 6_{c_2} p^2 q^4$$

Given 9P(X=4) = P(X=2)

$$9 * 6_{c_4} p^4 q^2 = 6_{c_2} p^2 q^4$$

$$135 p^2 = 15 q^2$$

$$9 p^2 = q^2$$

$$9 p^2 - q^2 = 0$$

$$9 p^2 - (1 - p)^2 = 0$$

$$9p^2-(1+p^2-2p)=0$$

$$9p^2-1-p^2+2p=0$$

$$8p^2+2p-1=0$$

$$p=\frac{-2\pm\sqrt{4+32}}{16}$$

$$p=\frac{-2\pm 6}{16} \qquad =\frac{4}{16},\frac{-8}{16}$$

$$p=\frac{1}{4},\frac{-1}{2}$$

Since p cannot be negative, p=1/4.

4. Out of 800 families with 4 children each, how many families would be expected to have

(i) 2 boys and 2 girls

(ii) at least 1 boy

(iii) at most 2 girls and

(iv) children of both sexes.

Assume equal probabilities for boys and girls.

Solution

Considering each child is a trial, n=4. Assuming that birth of a boy is success, p = 1/2 and q = ½

Let X denote the number of successes (boys)

(i) P[2 boys and 2 girls] = P(X=2)

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

$$P(X=2)=4_{c_2}(\frac{1}{2})^2(\frac{1}{2})^{4-2}$$

$$= 6(\frac{1}{2})^4 = \frac{3}{8}$$

Therefore number of families having 2 boys and 2 girls=N[P(X=2)]

= 800(3/8) = 100 * 3

= 300

(ii) P[at least 1 boy] = P[X≥1]

= P[X=1] + P[X=2] + P[X=3] + P[X=4]

=1- P[X=0]

$$P(X=0) = 4_{c_0}(\frac{1}{2})^0(\frac{1}{2})^{4-0}$$

$$1-P(X=0) = 1-(\frac{1}{2})^4 = \frac{15}{16}$$

Therefore number of families having at least 1 boy = N [1-(P(X=0)]

= 800 (15/16) = 750

(iii) P(at most 2 girls)= P(exactly 0 girl, 1 girl or 2 girls)

= P[X=4, X=3, X=2]

= 1-[P(X=0) + P(X=1)]

$$= 1-\left[4_{c_0}(\frac{1}{2})^0(\frac{1}{2})^{4-0} + 4_{c_1}(\frac{1}{2})^1(\frac{1}{2})^{4-1}\right]$$

$$= 1-\left[(\frac{1}{2})^4 + 4(\frac{1}{2})^4\right] = 1-(\frac{1}{16}+\frac{4}{16}) = 1-\frac{5}{16}$$

$$= \frac{11}{16}$$

Therefore number of families having at most 2 girls = N[P(X≥2)]

= 800 (11/16) = 550

(iv) P[children of both sexes] = 1 – P[children of same sex]

= 1 –[P(all are boys) + P(all are girls)]

= 1- [P(X=4) + P(X=0)]

$$=1-\left[4_{c_4}(\frac{1}{2})^4+4_{c_0}(\frac{1}{2})^0(\frac{1}{2})^4\right]=1-[(\frac{1}{2})^4+(\frac{1}{2})^4]$$

= 1- 2/16 = 7/8

Therefore number of families having children of both sexes = 800 * 7/8

= 700

5. An irregular 6 faced die is such that the probability that it gives 3 even numbers in 5 throws is twice the probability that it gives 2 even numbers in 5 throws. How many sets of exactly 5 trials can be expected to give no even number out of 2500 sets.

Solution

Let the probability of getting an even number with the unfair die bep .

Let X denote the number of even numbers obtained in 5 trials (throws)

Given: P(X=3) = 2 * P(X=2)

$$5_{c_3}\,p^3q^2=2*5_{c_2}\,p^2q^3$$

p = 2q

p = 2(1-p)

3p = 2

P = 2/3

q=1-p = 1/3

Now P[getting no even number] = P[X=0]

$$5_{c_0}\,p^0q^5=(\frac{1}{3})^5=\frac{1}{243}$$

Therefore number of sets having no success (even number) out of N sets = N [P(X=0)]

= 2500 * 1/243

= 10 nearly

7.. Assuming that half of the population is vegetarian and that 100 investigators each take 10 individuals to see whether they are vegetarians, how many would you expect to report that 3 people or less were vegetarians?

Solution

n=10, p=1/2, q=1/2

$$P(X = x) = n_{c_x} p^x q^{n-x}$$

$$= 10_{c_x} (\frac{1}{2})^x (\frac{1}{2})^{10-x}$$

$$= 10_{c_x} (\frac{1}{2})^{10}$$

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

$$= 10_{c_0} (\frac{1}{2})^{10} + 10_{c_1} (\frac{1}{2})^{10} + 10_{c_2} (\frac{1}{2})^{10} + 10_{c_3} (\frac{1}{2})^{10}$$

$$= (\frac{1}{2})^{10} [1 + 10 + 45 + 120] \backslash$$

$$= (\frac{1}{2})^{10} [176] = \frac{176}{1024} = 0.1718$$

Among 100 investigators, the number of investigators who report that 3 or less were consumers

=100 * 0.1718

=17 investigators

14. A factory produces 10 articles daily. It may be assumed that there is a constant probability p= 0.1 of producing a defective article. Before the articles are stored, they are inspected and the defective ones are set aside. Suppose that there is a constant probability r = 0.1, that a defective article is misclassified. If X denote the number of articles classified as defective at the end of a production day, find a) P(X=3) and b) P(X>3)

Solution

Let X be the random variable represented by the number of articles which are defective.

P[a defective article is classified as defective] = P(an article produced is defective) *P(it is classified as defective)

= 0.1 *0.9

p = 0.09

q = 1 – p = 0.91

n= 10

$$P(X = x) = n_{c_x} p^x q^{n-x}$$

$$= 10_{c_x} (0.09)^x (0.91)^{10-x}$$

$$P(X = 3) = 10_{c_3} (0.09)^3 (0.91)^7$$

$$= 0.0452$$

$$P(X > 3) = 1 - P(X \le 3)$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

$$= 1 - [10_{c_0} (0.09)^0 (0.91)^{10} + 10_{c_1} (0.09)^1 (0.91)^9 + 10_{c_2} (0.09)^2 (0.91)^8 + 10_{c_3} (0.09)^3 (0.91)^7]$$

=0.0089

POISSON DISTRIBUTION

Definition

If X is a discrete random variable that assumes only non-negative values such that its probability mass function is given by

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \text{ x} = 0,1,2,3,\ldots \quad \text{where } \lambda > 0$$
$$= 0, \text{otherwise}$$

then X is said to follow Poisson distribution with the parameter λ .

Poisson Distribution is a Limiting case of Binomial Distribution

Suppose in a binomial distribution,

1. The number of trails n is indefinitely large, i.e., $n \to \infty$.

2. The probability of success p for each trail is very small, i.e, $p \to 0$.

3. $np(=\lambda)$ is finite and $p = \frac{\lambda}{n}$, $q = 1 - p = 1 - \frac{\lambda}{n}$ where λ is a positive constant.

Mean of the Poisson distribution

$$\text{Mean} = E(X) = \sum_{x=0}^{\infty} xP(X = x)$$

$$= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{x\lambda^{x-1}}{x!}$$

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

$$\text{Mean} = \lambda e^{-\lambda} e^{\lambda}$$
$$= \lambda$$

Variance of Poisson distribution

$$\text{Var(X)} = E(X^2) - [E(X)]^2$$

$$[E(X)]^2 = \sum_{x=0}^{\infty} x^2 p(x)$$

$$= \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!}$$

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

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

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

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

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

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

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

$$E(X^2) = \lambda^2 + \lambda$$

$$\text{Var(X)} = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Therefore variance of the poisson distribution is λ

Examples:

1.If X is a Poisson variate such that P(X=1)=3/10 and P(X=2)=1/5. Find P(X=0) and P(X=3)

Solution

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

$$P(X = 1) = \frac{e^{-\lambda}\lambda^1}{1!} = \frac{3}{10} \qquad (1)$$

$$P(X = 2) = \frac{e^{-\lambda}\lambda^2}{2!} = \frac{1}{5} \qquad (2)$$

$$\frac{(2)}{(1)} \quad \Rightarrow \quad \frac{\frac{e^{-\lambda}\lambda^2}{2!}}{\frac{e^{-\lambda}\lambda}{1!}} = \frac{\frac{1}{5}}{\frac{3}{10}}$$

$$\frac{(2)}{(1)} \quad \Rightarrow \quad \frac{\lambda}{2} = \frac{2}{3} \quad \Rightarrow \lambda = \frac{4}{3}$$

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-\frac{4}{3}}\left(\frac{4}{3}\right)^{x}}{x!}$$

$$P(X=0)=\frac{e^{-\frac{4}{3}}\left(\frac{4}{3}\right)^{0}}{0!}=e^{-\frac{4}{3}}=0.2637$$

$$P(X=3)=\frac{e^{-\frac{4}{3}}\left(\frac{4}{3}\right)^{3}}{3!}=0.1047$$

2. In a certain factory producing razor blades, there is a small chance 1/500 for any blade to be defective. The blades are supplied in packets of 10.Use Poisson distribution to calculate the approximate number of packets containing

(i) no defective blade

(ii) at least 1 defective blade and

(iii) at most 1 defective blade in a consignment of 10,000 packets.

Solution

Given p=1/500 and n=10

Let X be the number of defectives in a packet

λ=np=10/500=1/50=0.02

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-0.02}(0.02)^{x}}{x!}$$

i) No defective blade : P(X=0)

$$=\frac{e^{-0.02}(0.02)^{0}}{0!}=0.9802$$

Therefore the number of pockets containing no defective razor= 10000 * 0.9802

= 9802

ii) At least 1 defective= P(X≥1)

= 1- P(X<1)

= 1-P(X=0)

= 1 – 0.9802 = 0.0198

Therefore the number of packets containing at least one defective= 10000 * 0.0198

= 198

iii) At most 1 defective = P(X≤1)

=P(X=0) + P(X=1)

$$= \frac{e^{-0.02}}{0!} + = \frac{e^{-0.02}(0.02)}{1!}$$

$$= 0.0198 + e^{-0.02}(0.02)$$

$$= 0.9997$$

Therefore the number of packets containing at most 1 defective blade = 10000 * 0.9997

= 9997

3. An insurance company has discovered that only about 0.1% of the population is involved in a certain type of accident each ear. If its 10000 policy holders were randomly selected from the population, what is the problem that not more than 5 of its clients are involved in such an accident next year?

Solution

Given p= 0.1% = 0.1/100 = 0.001

n= 10000

Mean $\lambda = np = 10000 * 0.001 = 10$

Let X be a random variable of number of clients involved in accident

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-10}(10)^x}{x!}$$

$$P(X \le 5) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$

$$P(X \le 5) = \frac{e^{-10}(10)^0}{0!} + \frac{e^{-10}(10)^1}{1!} + \frac{e^{-10}(10)^2}{2!} + \frac{e^{-10}(10)^3}{3!} + \frac{e^{-10}(10)^4}{4!} + \frac{e^{-10}(10)^5}{5!}$$

$$= e^{-10}\left\{1+\frac{10}{1}+\frac{100}{2}+\frac{1000}{6}+\frac{10000}{24}+\frac{100000}{120}\right\}$$

$$= 0.0671$$

4. In a given city 4% of all licenced drivers will be involved in at least 1 road accident in any given year. Determine the probability that among 150 licenced drivers ran only chosen in this city

i) only 5 will be involved in atleast 1 accident in any given year and

ii)at most 3 will be involved in atleast 1 accident in any given year.

Solution

$$\lambda = np = 100 \times \frac{4}{100} = 6$$

i) $$P(X=5) = \frac{e^{-6}6^5}{5!} = 0.1606$$

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

$$= e^{-6} + \frac{e^{-6}6}{1!} + \frac{e^{-6}6^2}{2!} + \frac{e^{-6}6^3}{31} = 0.1512$$

7. Messages arrive at a switch board in a Poisson manner at an average rate of six per hour. Find the probability for each of the following events

(i) Exactly two messages arrive within one hour

(ii) No message arrives with in one hour

(iii) at least three messages arrive within one hour

Solution

Mean $\lambda = 6$ per hour

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-6}6^x}{x!}$$

$$P(X=2) = \frac{e^{-6}6^2}{2!} = 0.0446$$

$$P(X=0)=\frac{e^{-6}6^{0}}{0!}=0.0025$$

$$P(X\geq 3)=1-P(X<3)=1-[P(X=0)+P(X=1)+P(X=2)]$$

$$=1-e^{-6}(1+6+18)=0.9380$$

8. A car hire firm has 2 cars which it hires out day by day. The number of demands for a car on each day follows a Poisson distribution with mean 1.5. Calculate the proportion of days on which

i) neither car is used

ii) some demand is not fulfilled

Solution

Let X be random variable representing the number of demands for cars:

P(x demands in a day)= $P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}$

Given: $\lambda = 1.5$

Now $P(X=x)=\frac{e^{-1.5}(1.5)^{x}}{x!}$

i) the proportion of days on which neither car is used

$$P(X=0)=\frac{e^{-1.5}1.5^{0}}{0!}=e^{-1.5}=0.2231$$

ii) The proportion of days on which some demand is refused

The demand is refused when x is more than 2

$$P(X>2)=1-[P(X\leq 2)]$$

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

$$=1-[\frac{e^{-1.5}(1.5)^{0}}{0!}+\frac{e^{-1.5}(1.5)^{1}}{1!}+\frac{e^{-1.5}(1.5)^{2}}{2!}]$$

$$=0.19126$$

9. The proofs of a 500 page book contains 500 misprints. Find the probability that there are at least 4 misprints in a randomly chosen page.

Solution

Total number of mistakes= 500

Total number of pages= 500

The average number of mistake per page is 1. $\lambda = 1$

Let X be a random variable of number of mistakes in a page.

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

$$\text{P(at least 4 mistakes)} = P(X \geq 4)$$

$$= 1 - P(X < 4)$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

$$= 1 - \left\{\frac{e^{-1}}{0!} + \frac{e^{-1}}{1!} + \frac{e^{-1}}{2!} + \frac{e^{-1}}{3!}\right\}$$

$$= 1 - e^{-1}\left\{1 + 1 + \frac{1}{2} + \frac{1}{6}\right\}$$

$$= 0.0180$$

NORMAL DISTRIBUTION

Definition

A normal distribution is a continuous distribution given by $y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ where X is a continuous normal variate distributed with density function

$f(\lambda) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ with mean μ and standard deviation σ.

Deviation of the distribution

When mean has been taken at the origin but if however another point is taken as the origin such that the excess of the mean over the arbitrary origin is m then

$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$ is the standard form of the normal curve with origin at (m,0).

Area under the normal curve is unity.

Characteristics of the Normal Distribution

The diagram of a normal distribution is given below. It is called normal curve.

Properties of the Normal Distribution

1. The normal distribution is a symmetrical distribution and the graph of the normal distribution is bell shaped.

2. The curve has a single peak point (i.e.,) the distribution is unimodal

3. The mean of the normal distribution lies at the centre of normal curve.

4.Because of the symmetry of the normal curve, the median and mode are also at the centre of the normal curve. Hence in a normal distribution the mean, median and mode coincide.

5. The tails of the normal distribution extend indefinitely and never touch the horizontal axes. That is we say that the normal curve approaches approximately from either side of its horizontal axes.

6. The normal distribution is a two parameter probability distribution. The parameters mean and standard deviation (μ,σ) completely determine the distribution.

7. Area property:

In a normal distribution about 67% of the observations will lie between mean $\pm$ S.D i.e., $(\mu \pm \sigma)$. About 95% of the observations lie between mean $\pm$ 2S.D (i.e., $\mu \pm 2\sigma$). About 99% of the observation will lie between mean $\pm$ 3S.D i.e.,$(\mu \pm 3\sigma)$.

Standard Normal Probability Distribution

If X is a normally distributed random variable, μ and σ are respectively its mean and standard deviation, then $Z = \frac{X - \mu}{\sigma}$ is called standard normal random variable.

Normal table

Special table called table of areas under normal curve is available to determine probabilities that the random variable lies in a given range of values of the variables. Using the table, we can determine the probability for X, taking a value less than x (X<x) and also for a given probability we determine the value x such that X < x

Additive property of Normal Distribution:

If $X_1, X_2, \ldots, X_n$ are independent normal variates with parameters $(m_1,\sigma_1),(m_2,\sigma_2)$

$\ldots,(m_n, \sigma_n)$ respectively then $X_1+X_2+\ldots..+ X_n$ is also a normal variate with parameter (m, σ)

Where $m=m_1+m_2+\ldots.+m_n$ and $\sigma^2= \sigma_1^2+ \sigma_2^2+\ldots\ldots..+ \sigma_n^2$.

Examples

1) X is a normal variate with mean 30 and standard deviation 5. Find the probability that

i) $26 \le X \le 40$; ii) $X \ge 45$ iii) $|X-30| > 5$.

Solution

Given $\mu = 30; \sigma = 5$

$$z = \frac{X-\mu}{\sigma}$$

i) when X= 26, $z = \frac{26-30}{5} = -0.8$

when X=40, $z = \frac{40-30}{5} = 2$

$$\therefore P[26 \le X \le 40] = P[-0.8 \le z \le 2]$$

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

$$= P[0 \le z \le 0] + P.8[0 \le z \le 2]$$

=0.2881+0.4772

=0.7653.

ii) when X=45, $z = \frac{45-30}{5} = 3$

$$\therefore P(X \ge 45) = P(z \ge 3)$$

=0.5- $P(0 \le z \le 3)$

=0.5-0.4987=0.0013.

iii) To find $P(|X-30|>5)$

$P(|X-30| \le 5) = P(25 \le X \le 35)$

When X=25, $z = \frac{25-30}{5} = -1$

When X=35, $z = \frac{35-30}{5} = 1$

$= 2P(0 \le z \le 1)$

=2(0.3413)

=0.6826.

$\therefore P(|X-30|>5) = 1 - P(|X-30| \le 5)$

=1-0.6826

=0.3174

2. A normal distribution has mean $\mu = 20$ and standard deviation $\sigma = 10$ .Find $P(15 \le X \le 40)$.

Solution

Given $\mu = 20$ and $\sigma = 10$

We know that $z = \frac{X-\mu}{\sigma}$.

When $X = 15$, $z = \frac{15-20}{10} = -0.5$ and

When =40, $z = \frac{40-20}{10} = 2$

$P(-0.5 \le z \le 2) = P(0 \le z \le 2) + P(0 \le z \le 0.5)$

=0.4772+0.1915

=0.6687.

3. The average seasonal rainfall in a place is 16 inches with a standard deviation of 4 inches. What is the probability that in a year the rainfall in that place will be between 20 and 24 inches?

Solution

$$z = \frac{X - \mu}{\sigma}$$

When X=20, $z = \frac{20-16}{4} = 1$

When X =24, $z = \frac{24-16}{4} = 2$

$\therefore P(20 < X < 24) = P(1 < z < 2)$

$= P(0 < z < 2) - P(0 < z < 1)$

=0.4772-0.3413

=0.1359.

Note

E(aX+bY)=aE(X)+bE(Y)

Var(aX+bY)=a^2V(X)+b^2V(Y)

Var(a)=0

E(a)=a

4. X is a normal variate with mean 1 and variance 4. Y is another normal variate independent of X with mean 2 and variance 3. What is the distribution of X+2Y?

Solution

Since X and Y are independent normal variates, X+2Y will also be a normal variate by the additive property and

Mean of X+2Y= E(X+2Y)=E(X)+2E(Y)

=1+2(2)=5

Variance of X+2Y=V(X+2Y)=V(X)+2^2V(Y)

=4+4(3)=16.

∴ X+2Y will follow normal with mean 5 and variance 16.

5. The saving bank account of a customer showed an average balance of Rs.150 and a standard deviation of Rs.50. Assuming that the account balances are normally distributed.

1. What percentage of account is over Rs. 200?
2. What percentage of account is between Rs.120 and Rs.170?
3. What percentage of account is less than Rs.75?

Solution

1) To find $P(X \geq 200)$

We know that $z = \frac{X - \mu}{\sigma}$

When X=200, $z = \frac{200 - 150}{50} = 1$

$P(X \geq 200) = P(z \geq 1) = 0.5 - P(0 < z < 1)$

=0.5-0.3413

=0.1587.

∴ Percentage of account is over Rs. 200 is 15.87%.

2. To find $P(120 < X < 170)$

When X= 120, $z = \frac{120 - 150}{50} = -0.6$

When X=170, $z = \frac{170 - 150}{50} = 0.4$

$\therefore P(120 < X < 170) = P(-0.6 < z < 0.4)$

$= P(0 < z < 0.6) + P(0 < z < 0.4)$

=0.2257+0.1554=0.3811

Therefore, percentage of account between Rs.120 and Rs.170 is 0.3811(100)=38.11.

3. To find $P(X < 75)$

When X=75, $z = \frac{75-150}{50} = -1.5$

$\therefore P(X < 75) = P(z < -1.5)$

$= 0.5 - P(0 < z < 1.5)$

=0.5-0.4322=0.0668.

Therefore, percentage of account is less than Rs.75 is 6.68%

6. The mean yield for one-acre plot is 662 kilos with standard deviation 32 kilos. Assuming normal distribution, how many one-acre plots in a patch of 1000 plots would you expect to have yield over 700 kilos below 650 kilos.

Solution

Given $\mu = 662, \sigma = 32$

$z = \frac{X - \mu}{\sigma} = \frac{X - 662}{32}$

When X=700, $z = \frac{700 - 662}{32}$ =1.19

When X=650, $z = \frac{650 - 662}{32} = -0.375 = -0.38$

$P[X > 700] = P(z > 1,19)$

$= 0.5 - P(0 \le z < 0.38) = 0.352$

Therefore, the number of plots have yield below 650 kilos=352.

UNIT II

Descriptive Statistics

Measures of Central Tendency

Types of Data:

(i) Individual observations
(ii) Discrete series
(iii) Continuous series

Example: (i)45, 56,78,97,…….,90

(ii)X (height of the student) f(no. of student)

155	7
156	4
157	2
158	5
159	1
160	1

	20

(iii)Marks X No.of students

0-10	0
10-20	0

20-30	1
30-40	1
40-50	2
50-60	1
60-70	5
70-80	4
80-90	4
90-100	1
	20

Measures of Central Tendency:

Individual Observations

Mean = $\frac{\sum x}{n}$

Mean = $A + \frac{\sum d}{n}$ where d=x-A, A-Assumed mean(Individual Observations)

Median = $\frac{n+1}{2}$ th item(Individual Observations)

Mode = the item which is occurred more number of times.

Discrete Series:

Mean = $\frac{\sum fx}{N}$ where $N = \sum f$

Mean = $A + \frac{\sum fd}{N}$ where d=x-A, A-Assumed mean

Mean = $A + \frac{\sum fd}{N} \times i$ where $d = \frac{x - A}{i}$, A-Assumed mean , i= class interval

Median = $\frac{N+1}{2}$ th item

Mode = the item which is occurred more number of times.

Continuous Series:

Mean = $\frac{\sum fm}{N}$ where $N = \sum f$

Mean = $A + \frac{\sum fd}{N}$ where d=m-A, A-Assumed mean

Mean = $A + \frac{\sum fd}{N} \times i$ where $d = \frac{m - A}{i}$, A-Assumed mean , i= class interval

Median = $\frac{N}{2}$th item

Median = $L + \frac{\frac{N}{2} - cf}{f} \times i$

Where L- Lower limit of the median class

$N = \sum f$

Cf-cumulative frequency preceding the median class

f- frequency of the median class

i-class interval

Mode = $M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$

Where L – lower limit of the modal class

f_0 -frequency of the class preceding the modal class

f_1 - frequency of the modal class

f_2 - frequency of the class succeeding the modal class

i-class interval

Empirical relation:

Mean - Mode =3 (Mean – Median)

(OR)

Mode = Mean - 3 Mean + 3 Median

(OR)

Mode = 3 Median - 2 Mean

Problems :

1.Find the AM , median and mode of the following set of observations:
25,32,28,34,24,31,36; 27,29,30.

Mean= $\frac{\sum x}{n}$

=(25+32+28+34+24+31+36+ 27+29+30)/10=296/10

=29.6

24,25,27,28,29,30,31,32,34,36

Median =(n+1)/2th item

=(10+1)/2 th item=5.5^{th} item

5.5^{th} item =(5^{th} item+6^{th} item)/2=(29+30)/2=29.5

There is no mode.

2. Find the mode of following data: **2,3,2,1,3,2,3,3,2,1,3,3,3,2,2,1,1,3,3,3**

Mode =3 [since 3 come more number of times]

3. Find the mean, median and mode.

x	f
155	30
156	20
157	5
158	5
159	10
160	15
161	5
162	10

Solution:

x	f	fx	cf(cumulative frequency)
155(mode)	30	4650	30
156	20	3120	50
157(Median)	5	785	55
158	5	790	60
159	10	1590	70
160	15	2400	85
161	5	805	90
162	10	1620	100
	$N = \sum f$	$\sum fx$	

	=100	=15760	

N=100

Mean= $\frac{\sum fx}{N}$

=15760/100=157.60

$\overline{X} = 157.60$

Median = (N+1)/2 th item=(100+1)/2th item= 50.5^{th} item

Median =157

Mode =155

4.Find the mean, median and mode for the following data:

Class(x)	frequency(f)
0-10	20
10-20	5
20-30	3
30-40	8
40-50	10
50-60	35
60-70	10
70-80	4
80-90	3
90-100	2
	$N = \sum f$ =100

Solution:

Class(x)	**m**	**frequency(f)**	**fm**	**cf**
0-10	5	20	100	20
10-20	15	5	75	25
20-30	25	3	75	28
30-40	35	8	280	36
40-50	45	10f0	450	46
50-60(median, Mode)	**55**	**35f1**	**1925**	**81**
60-70	65	10f2	650	91
70-80	75	4	300	95
80-90	85	3	255	98
90-100	95	2	190	100
		$N = \sum f$ **=100**	$\sum fm$ **=4300**	

N=100

Mean = $\frac{\sum fm}{N}$ where $N = \sum f$

=4300/100=43

Median =(N/2)th item = (100/2)th item =50^{th} item

50^{th} item is 50-60 (the median class)

Median $= L + \dfrac{\frac{N}{2} - cf}{f} \times i$

$= 50 + \dfrac{50-46}{35} \times 10$

=50+1.1428=51.1428

Modal class is 50-60

$M_0 = L + \dfrac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$

$= 50 + \dfrac{35-10}{2(35)-10-10} \times 10$

=50+5=55

Another method:

Class(x)	m	frequency(f)	$d = \dfrac{m-55}{10}$	fd
0-10	5	20	-5	-100
10-20	15	5	-4	-20
20-30	25	3	-3	-9
30-40	35	8	-2	-16
40-50	45	10	-1	-10
50-60	55	35	0	0
60-70	65	10	1	10
70-80	75	4	2	8
80-90	85	3	3	9
90-100	95	2	4	8
		$N = \sum f$ =100		$\sum fd$ =-120

Mean = $A + \dfrac{\sum fd}{N} \times i$ where $d = \dfrac{m-A}{i}$, A-Assumed mean , i= class interval

A=55, i=10

Mean= $55 + \dfrac{-120}{100} \times 10$

=55-12=43

MEASURES OF DISPERSION

There are five measures of dispersion:

Range,

Inter-quartile range or Quartile Deviation,

Mean deviation,

Standard Deviation,

and Lorenz curve.

Among them, the first four are mathematical methods and the last one is the graphical method.

Range:

Range = L-S

L- Largest values ; S- Smallest value

Coefficient of Range = $\frac{L-S}{L+S}$

Problems:

Example 1: Find the range for the following sets of data:

5, 15, 15, 05, 15,05,15,15,15,15

Range = 15-5=10

Coeff.of range= (L-S)/(L+S)

=10/20=1/2

2. Find the range for the following sets of data:

8 , 7, 15, 11, 12, 5, 13, 11 ,15, 9

Range = 15-5=10

Coeff.of range= (L-S)/(L+S)

=10/20=1/2

Example 2: Calculate the coefficient of range separately for the two sets of data

given below:

Set 1	8	10	20	9	15	10	13	28
Set 2	30	35	42	50	32	49	39	33

Solution: It can be seen that the range in both the sets of data is the same:

Set 1 28-8=20

Set 2 50-30=20

Coefficient of range in Set 1 is:

$$\frac{28-8}{28+8} = 0.55$$

Coefficient of range in set 2 is:

$$\frac{50-30}{50+30} = 0.25$$

3.

3. Compute the range

X	**f**
158	15
159	20
160	32
161	35
162	33
163	22
164	20
165	10
166	8
	N=195

Range = 166-158=8

Coeff. Of Range = (166-158)/(166+158)=0.0246

4: Find the range for the following frequency distribution:

Size of Item	Frequency
20- 40	7
40- 60	11
60- 80	30
80-100	17
100-120	5
Total	**70**

Here, the upper limit of the highest class is 120

and the lower limit of the lowest class is 20.

Hence, the range is 120 - 20 = 100.

The coefficient of range is calculated by the formula: (L-S)/ (L+S)

Quartile Deviation :

$$QD = \frac{Q_3 - Q_1}{2}$$

Q_1- First quartile

Q_3-3rd quartile

Coefficient of quartile deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

To find Q1:

Q1 = $\frac{N+1}{4}$ **th item (Individual observations)**

Q1 = $\frac{N+1}{4}$ **th item (Discrete series)**

Q1 = $\frac{N}{4}$ **th item (continuous series)**

$$Q1 = L + \frac{\frac{N}{4} - cf}{f} \times i$$

To find Q3:

Q3 = $\frac{3(N+1)}{4}$ **th item (Individual observations)**

Q3= $\frac{3(N+1)}{4}$ **th item (Discrete series)**

Q3 = $\frac{3N}{4}$ **th item (continuous series)**

$$Q3 = L + \frac{\frac{3N}{4} - cf}{f} \times i$$

1. Find the Quartile deviation and the coefficient of QD 3, 8, 6, 10, 12, 9, 11, 10, 12, 7

Solution:

3, 6 , 7, 8,9,10,10,11,12,12

N=10

Q1= 2.75th item

Q1= 2nd item + 0.75(3rd item – 2nd item)

= 6+0.75(7-6)=6.75

Q3= 3(10+1)/4 th item = 8.25th item

Q3=8th item + 0.25(9th item -8th item)

= 11+0.25(12-11)=11.25

QD=(11.25-6.75)/2=2.25

QD= (Q3-Q1)/2

Coefficient of QD=0.25

2. **Compute the Quartile deviation and its relative measure.**

X	**f**
158	15
159	20
160	32
161	35
162	33
163	22
164	20
165	10
166	8
	N=195

Solution:

X	**f**	**cf**
158	15	15
159	20	35
160Q1	32	67
161	35	102
162	33	135
163Q3	22	157
164	20	177
165	10	187
166	8	195
	N=195	

Q1 = $\frac{N+1}{4}$ **th item (Discrete series)**

=(195+1)/4th item=49 th item

Q1=160

Q3= $\frac{3(N+1)}{4}$ **th item (Discrete series)**

=3(195+1)/4=147th item

Q3=163

QD=(Q3-Q1)/2

=(163-160)/2=1.5

CQD=0.0092

3. Find the quartile deviation and its relative measure.

Class(x)	**frequency(f)**
0-10	20
10-20	5
20-30	3
30-40	8
40-50	10
50-60	**35**
60-70	10
70-80	4
80-90	3
90-100	2
	$N = \sum f$ **=100**

Solution:

Class(x)	**frequency(f)**	**cf**
0-10	20	20
10-20	5	25
20-30 Q1	3	28
30-40	8	36
40-50	10	46
50-60Q3	35	**81**
60-70	10	91
70-80	4	95
80-90	3	98
90-100	2	100
	$N = \sum f$ **=100**	

Q1=(100/4)th item =25th item

Q1 lies between 20-30

$$Q1 = L + \frac{\frac{N}{4} - cf}{f} \times i$$

=20 +(25-25)/3 x10
=20
Q3= 3(100)/4th item
= 75 th item

Q3 lies between 50-60

$$Q3 = L + \frac{\frac{3N}{4} - cf}{f} \times i$$

=50+(75-46)/35 X10 = 50+0.829X10
=58.29

QD= (Q3-Q1)/2
=(58.29-20)/2
=19.15
CQD=(58.29-20)/ (58.29+20)
= 0.4890
Mean Deviation:

Mean Deviation about mean:

$$MD = \frac{\sum |D|}{N}$$ **where** $D = x - \bar{x}$ **(Individual Observations)**

(OR)

Mean Deviation about median:

$$MD = \frac{\sum |D|}{N}$$ **where** $D = x - \text{Median}$ **(Individual Observations)**

Mean Deviation about mean:

$$MD = \frac{\sum f|D|}{N}$$ **where** $D = x - \bar{x}$ **(Discrete series)**

(OR)

Mean Deviation about median:

$$MD = \frac{\sum f|D|}{N}$$ **where** $D = x - \text{Median}$ **(Discrete series)**

Mean Deviation about mean:

$$MD = \frac{\sum f|D|}{N}$$ **where** $D = m - \bar{x}$ **(Continuous series)**

(OR)

Mean Deviation about median:

$$MD = \frac{\sum f|D|}{N}$$ **where** $D = m - \text{Median}$ **(Continuous series)**

Coefficient Mean Deviation = MD/mean

Coefficient Mean Deviation = MD/median

PROBLEMS:

1. Find the MD of the set of numbers 3, 8, 6, 10, 12, 9, 11, 10, 12, 7
   Soln:

X	|D|=|x-8.8|
3	5.8
8	0.8
6	2.8
10	1.2
12	3.2
9	0.2
11	2.2
10	1.2
12	3.2
7	1.8
Total= $\sum X = 88$	$\sum \lvert D \rvert = 22.4$

N=10

Mean = $\bar{X}$ =88/10

$\bar{X}$ =8.8

Mean deviation about mean= $\frac{\sum |D|}{N}$

= 22.4/10=2.24

2. Compute the Mean deviation about median.

Solution:

X	**f**	**cf**	**|D|=|x-161|**	**f|D|**
158	15	15	3	45
159	20	35	2	40
160	32	67	1	32
161 Median	35	102	0	0
162	33	135	1	33
163	22	157	2	44
164	20	177	3	60
165	10	187	4	40
166	8	195	5	40
	N=195			$\sum \mathbf{f\lvert D\rvert = 334}$

Median = (N+1)/2 th item=(195+1)/2th item= 98^{th} item

Median =161

Mean Deviation about median:

$MD = \frac{\sum f|D|}{N}$ where $D = x - \text{Median}$ (Discrete series)

Mean deviation = 334/195=1.712

Coefficient of Mean deviation = MD/Median = 1.712/161=0.010

3. Find the Mean deviation for the following data:

Size of Item	Frequency
2-4	20
4-6	40
6-8	30
8-10	10

Solution:

X	**m**	**f**	**fm**	**|D|=|m-5.6|**	**f|D|**
2-4	3	20	60	2.6	52
4-6	5	40	200	0.6	24
6-8	7	30	210	1.4	42
8-10	9	10	90	3.4	34
		N=100	$\sum$fm **=560**		$\sum$f | D | **=152**

Mean= $\frac{\sum fm}{N}$

=560/100=5.6

Mean Deviation about mean:

$MD = \frac{\sum f|D|}{N}$ **where** $D = m - \bar{x}$ **(Continuous series)**

MD=152/100=1.52

CMD=MD/Mean

=1.52/5.6=0.271

Standard Deviation:

$\sigma = \sqrt{\frac{\sum x^2}{n}}$ **where** $x = X - \bar{X}$ **(Individual Observations)**

$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$ **where** $d = X - A$**; A- Assumed mean(Individual Observations)**

$\sigma = \sqrt{\frac{\sum fx^2}{n}}$ **where** $x = X - \bar{X}$ **(Discrete series)**

$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$ **where** $d = X - A$**; A- Assumed mean(Discrete series)**

$\sigma = \sqrt{\frac{\sum fx^2}{n}}$ **where** $x = m - \bar{X}$ **(Continuous series)**

$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times i$ **where** $d = \frac{m - A}{i}$**; A- Assumed mean(Continuous series)**

1. Find the Standard deviation of the set of numbers 3, 8, 6, 10, 12, 9, 11, 10, 12, 7
   Soln:

X	$x = X - 8.8$	x^2
3	-5.8	33.64
8	-0.8	0.64
6	-2.8	7.84
10	1.2	1.44
12	3.2	10.24
9	0.2	0.04
11	2.2	4.84
10	1.2	1.44
12	3.2	10.24
7	-1.8	3.24
Total= $\sum X$ =88		$\sum x^2$ =73.6

N=10

Mean = $\bar{X}$ =88/10

$\bar{X}$ =8.8

Standard deviation = $\sigma = \sqrt{\frac{\sum x^2}{n}}$ where $x = X - \bar{X}$

$$\sigma = \sqrt{\frac{73.6}{10}}$$

$$= \sqrt{7.36} = 2.712$$

2. Compute the Standard deviation.

X	**f**	fx	$x = X - 161.5128$	$\mathbf{x^2}$	$\mathbf{fx^2}$
158	15	2370	-3.5128	12.3398	185.0965
159	20	3180	-2.5128	6.3142	126.2833
160	32	5120	-1.5128	2.2886	73.23404
161	35	5635	-0.5128	0.2630	9.2037
162	33	5346	0.4872	0.2374	7.8330
163	22	3586	1.4872	2.2118	48.6588
164	20	3280	2.4872	6.1862	123.7233
165	10	1650	3.4872	12.1606	121.606
166	8	1328	4.4872	20.1350	100.6748
	N=195	31495			856.7055

Mean= $\frac{\sum fx}{N}$

= 31495/195=161.5128

$\sigma = \sqrt{\frac{\sum fx^2}{n}}$ **where** $x = X - \overline{X}$ **(Discrete series)**

$= \sigma = \sqrt{\frac{856.7055}{195}}$

=2.09603

3.Find the standard deviation for the following data:

Size of Item	**Frequency**
2-4	20
4-6	40
6-8	30
8-10	10

Coefficient of Variation:

$$CV = \frac{\sigma}{\overline{X}} \times 100$$

4. Find the Range, Quartile deviation , Mean deviation and standard deviation for the following data:

Size of Item	Frequency
2-4	20
4-6	40
6-8	30
8-10	10

Solution:

range =

10-2=8

Size of Item	Mid-points (m)	Frequency (f)	cf	fm	|D|=|m-5.6|	f |D|	d=(m-A)/2	fd	$f(d^2)$
2-4	3	20	20	60	2.6	52	-1	-20	20
4-6(Q1)	**5**	**40**	**60**	200	0.6	24	0	0	0
6-8Q3	**7**	**30**	**90**	210	1.4	42	1	30	30
8-10	9	10	100	90	3.4	34	2	20	40
	Total	**100**		**560**		**152**		**30**	**90**

Here N=100

$$MD = \frac{\sum f\,|D|}{N} \quad \textbf{where} \quad D = m - \bar{x}$$

Mean=5.6

$$MD = \frac{152}{100}$$

MD=1.52

Q_1=(N/4)th item

=25th item

Q_1 lies in 4-6

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times i$$

$$= 4 + \frac{25 - 20}{40} \times 2$$

Q_1=4.25

Q_3=(3N/4)th item

=75th item

Q_3 lies in 6-8

$$Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \times i$$

$$= 6 + \frac{75 - 60}{30} \times 2$$

Q_3=7

$$\textbf{QD} = \frac{Q_3 - Q_1}{2}$$

=(7-4.25)/2=1.375

Coefficient of QD=0.244

$$\sigma = \sqrt{\frac{\sum fx^2}{n}} \quad \textbf{where } d = m - \overline{X}$$

$$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times i \quad \textbf{where } d = \frac{m - A}{i}\textbf{; A- Assumed mean}$$

$$\sigma = \sqrt{\frac{90}{100} - \left(\frac{30}{100}\right)^2} \times 2$$

=1.8

CV=32.1

6. Compute Range, QD, MD and SD.

X	**f**	**cf**	**fx**	**|D|=|X-mean|**	**|D|=|X-median|**	**f|D(mean)|**	**f|D|(median)**	**x=X-mean**	**x^2**	**$f(x^2)$**
158	15	15	2370	3.5128	3	52.692	45	-3.5128	12.3398	185.0965
159	20	35	3180	2.5128	2	50.256	40	-2.5128	6.3142	126.2833
$Q_1$160	32	67	5120	1.5128	1	48.4096	32	-1.5128	2.2886	73.23404
M161	35	102	5635	0.5128	0	17.948	0	-0.5128	0.2630	9.2037
162	33	135	5346	0.4872	1	16.0776	33	0.4872	0.2374	7.8330
$Q_3$163	22	157	3586	1.4872	2	32.7184	44	1.4872	2.2118	48.6588
164	20	177	3280	2.4872	3	49.744	60	2.4872	6.1862	123.7233
165	10	187	1650	3.4872	4	34.872	40	3.4872	12.1606	121.606
166	8	195	1328	4.4872	5	35.8976	40	4.4872	20.1350	100.6748
	N=195		31495			338.6152	334			856.7055

Mean=161.5128

Median=(195+1)/2th item=98th item=161

Q1=(195+1)/4th item=49th item=160

Q3=3(195+1)/4th item=147th item=163

QD=1.5

CQD=0.0092

MD about mean=1.736

CMD about mean=1.736/161.5128=0.01075

MD about median=1.7128

CMD about median=1.7128/161=0.01063

$$SD=\sqrt{\frac{856.7055}{195}} = 2.096$$

CV=(2.096/161.5128)X100=1.2977

Range=8

7. The following are scores of two batsmen A and B in a series of innings:

A: 12 115 6 73 7 19 119 36 84 29

B: 47 12 16 42 4 51 37 48 13 0

Who is better scorer and who is more consistent?

Solution:

X	x=X-mean	x^2	Y	y=Y-mean	y^2
12	-38	1444	47	20	400
115	65	4225	12	-15	225
6	-44	1936	16	-11	121
73	23	529	42	15	225
7	-43	1849	4	-23	529
19	-31	961	51	24	576
119	69	4761	37	10	100
36	-14	196	48	21	441
84	34	1156	13	-14	196
29	-21	441	0	-27	729
total=500		17498	270		3542

N=10

$$\overline{X} = \frac{\sum x}{n}$$

=500/10=50

$$\overline{Y} = \frac{\sum y}{n}$$

270/10=27

$$\sigma_x = \sqrt{\frac{\sum x^2}{n}}$$

$$\sigma_x = \sqrt{\frac{17498}{10}} = \sqrt{1749.8}$$

$$\sigma_x = 41.8306$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{n}}$$

$$\sigma_y = \sqrt{\frac{3542}{10}} = \sqrt{354.2} = 18.846$$

CV for A:

$$cv = \frac{\sigma_x}{\bar{X}} \times 100$$

83.6612%

CV for B:

$$\textbf{CV=} \frac{\sigma_y}{\bar{Y}} \times 100$$

69.6%

Mean of A > Mean of B

Therefore A is the better player.

CV for A > CV for B

So here B is the consistent player.

MEASURES OF SKEWNESS

Skewness

Literal meaning of skewness is lack of symmetry. It measures the degree of departure of a distribution from symmetry and reveals the direction of scatterdness of the items.

A frequency distribution is said to be symmetrical when values of the variables equidistant from their mean have equal frequencies. If a frequency distribution is not symmetrical, it is said to be asymmetrical or skewed. Any deviation from symmetry is called skewness.

According to *Morris Humberg* Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution.

According to *Croxton & Cowden* When a series is not symmetrical it is said to be asymmetrical or skewed.

According to *Simpson & Kafka* Measures of skewness tell us the direction and the extent of skewness. In a symmetrical distribution the mean, median and mode are identical. The more we move away from the mode, the larger the asymmetry or skewness.

Symmetrical curve

The figure , given below, presents the shape of a symmetrical curve which is bell shaped having no skewness. The value of mean (M), median (M_d) and mode (M_o) for such a curve would be identical.

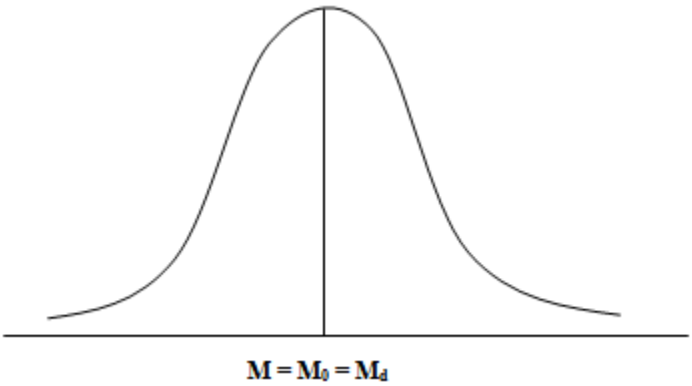


Fig. 5.1 Symmetrical distribution

In a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the centre point of the curve. For a symmetrical distribution Mean = Median = Mode.

Positively skewed curve

A positively skewed curve has a longer tail towards the higher values of X i.e. the frequency curve gradually slopes down towards the higher values of X. In a positively skewed distribution the mean is greater than the median and then mode and the median lies in between mean and mode. The frequencies are spread over a greater range of values on the high value end of the curve (the right hand side) as is clear from the Figure

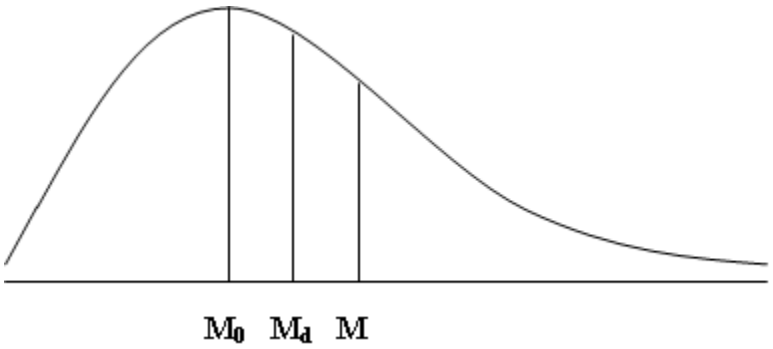


Fig. Positively skewed distribution

For a positively skewed distribution Mean > Median > Mode.

Negatively skewed curve

A negatively skewed curve has a longer tail towards the lower values of X i.e. the frequency curve gradually slopes down towards the lower values of X as shown in Figure.

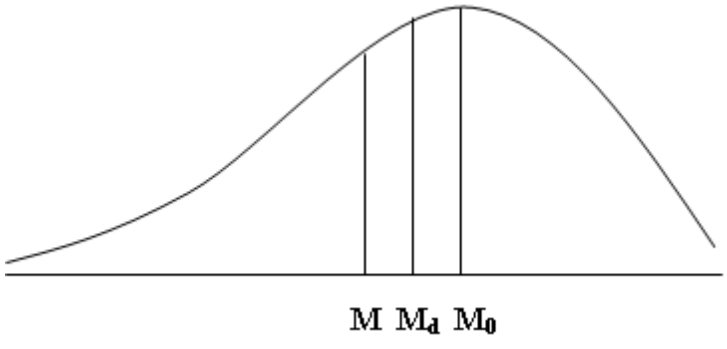


Fig. Negatively skewed distribution

In the negatively skewed distribution the mode is the maximum and mean is the least. The median lies in between mean and mode. The elongated tail in negatively skewed distribution is on the left hand side as would be clear from Figure. For a negatively skewed distribution,

Mean < Median < Mode.

Karl pearson's coefficient of skewness

The first coefficient of skewness as defined by Karl Pearson is

$$\text{Coefficient of skewness} = \frac{\text{Mean}-\text{Mode}}{\text{Std.deviation}} = \frac{M-M_0}{\sigma}$$

This measure is based on the fact that the mean and the mode are drawn widely apart. Skewness will be positive if mean > mode and negative if mean < mode. There is no limit to this measure in theory and this is a slight drawback. But in practice the value given by this formula is rarely very high and its value usually lies between -1 and +1.

It may also be written as $\frac{3(\text{Mean}-\text{Median})}{\sigma}$ as Mode = 3 Median - 2 Mean

This coefficient is a pure number without units since both numerator and denominator have the same dimensions. The value of this coefficient lies between -3 and +3.

Problems:

1. Calculate the Karl Pearson's coefficient of Skewness :
   3, 8, 6, 10, 12, 9, 11, 10, 12, 7

   Solution:
   Ascending order : 3, 6, 7,8, 9,10,10,11,12,12
   Mean= $\frac{\sum x}{n}$
   =88/10=8.8

We can't define Mode.

Median = (n+1)/2th item = (10+1)/2th item
=5.5 th item

Median = (5^{th} item+6^{th} item)/2=(9+10)/2=9.5

X	$x = X - 8.8$	x^2
3	-5.8	33.64
8	-0.8	0.64
6	-2.8	7.84
10	1.2	1.44
12	3.2	10.24
9	0.2	0.04

11	2.2	4.84
10	1.2	1.44
12	3.2	10.24
7	-1.8	3.24
Total= $\sum X = 88$		$\sum x^2 = 73.6$

Standard deviation $= \sigma = \sqrt{\frac{\sum x^2}{n}}$ where $x = X - \overline{X}$

$$\sigma = \sqrt{\frac{73.6}{10}}$$

$$= \sqrt{7.36} = 2.712$$

Karl Pearson's coefficient of skewness= $\frac{3(\overline{X} - \text{Median})}{\sigma}$

=3(8.8-9.5)/2.712=-0.77433

Negative skewed.

2. Find the Karl Pearson's coefficient of Skewness :

x	f
155	30
156	20
157	5
158	5
159	10
160	15
161	5
162	10

Solution:

X	f	fX	x=X-mean	x^2	f x^2
155(mode)	30	4650	-2.60	6.76	20.28
156	20	3120	-1.60	2.56	51.20
157	5	785	-0.60	0.36	1.8
158	5	790	0.40	0.16	0.8
159	10	1590	1.40	1.96	19.6
160	15	2400	2.40	5.76	86.4

161	5	805	3.40	11.56	57.8
162	10	1620	4.40	19.36	193.6
	$N = \sum f$ =100	$\sum fx$ =15760			$\sum f\ x^2$ =431.48

N=100

Mean= $\frac{\sum fx}{N}$

=15760/100=157.60

$\overline{X} = 157.60$

Mode =155

Standard deviation = $\sqrt{\frac{\sum fx^2}{N}}$ x=X-Mean

$= \sqrt{\frac{431.48}{100}} = 2.077$

Skewness= $\frac{\overline{X} - \text{Mode}}{\sigma}$

=(157.60-155)/2.077=1.252

Positively skewed.

3.Find coefficient of skewness for the following data:

Class(x)	**frequency(f)**
0-10	20
10-20	5
20-30	3
30-40	8
40-50	10
50-60	35
60-70	10
70-80	4
80-90	3
90-100	2
	$N = \sum f$ **=100**

Solution:

Class(x)	**m**	**frequency(f)**	**fm**	x=m-mean	x^2	$f\ x^2$
0-10	5	20	100	-38	1444	28880
10-20	15	5	75	-28	784	3920

20-30	25	3	75	-18	324	972
30-40	35	8	280	-8	64	512
40-50	45	10f0	450	2	4	40
50-60(Mode)	**55**	**35f1**	**1925**	**12**	**144**	**5040**
60-70	65	10f2	650	22	484	4840
70-80	75	4	300	32	1024	4096
80-90	85	3	255	42	1764	5292
90-100	95	2	190	52	2704	5408
		$N=\sum f$ **=100**	$\sum fm$ **=4300**			**59000**

N=100

Mean = $\frac{\sum fm}{N}$ where $N=\sum f$

=4300/100=43

Modal class is 50-60

$$M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 50 + \frac{35-10}{2(35)-10-10} \times 10$$

=50+5=55

Standard deviation= $\sqrt{\frac{\sum fx^2}{N}}$ x=m-Mean

$$= \sqrt{\frac{59000}{100}} = \sqrt{590}$$

=24.28

Skewness=(43-55)/24.28=-0.4942

Negatively skewed.

4.. Find the coefficient of skewness from the following data.

A=22.5

X	f	m	cf	d=(m-A)/5	fd	d^2	fd^2	
0-5	2	2.5	2	-4	-8	16	32	
5-10	5	7.5	7	-3	-15	9	45	
10-15	7	12.5	14	-2	-14	4	28	

Q1 15-20	13	17.5	27	-1	-13	1	13	
M 20-25	21	22.5	48	0	0	0	0	
Q3 25-30	16	27.5	64	1	16	1	16	
30-35	8	32.5	72	2	16	4	32	
35-40	3	37.5	75	3	9	9	27	
Total	75				-9		193	

Q1=15+ $\frac{18.75-14}{13}\times 5$ **=16.827**

Median= 20+ $\frac{37.5-27}{21}\times 5$ **=22.5**

Q3=27.57

Sk=(27.57+16.827-45)/(27.57-16.827)=-0.055

Negatively skewed.

UNIT III

Multivariate Analysis

Correlation and Regression analysis

Correlation Coefficient :

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

$$r = \frac{\sum x^2 \sum y^2}{\sum xy}$$

Where $x = X - \overline{X}$ and $y = Y - \overline{Y}$

$$r = \frac{N\sum dxdy - (\sum dx)(\sum dy)}{\sqrt{N\sum dx^{2-} (\sum dx)^2}\sqrt{N\sum dy^2 - (\sum dy)^2}}$$

Where dx=X-A

dy=Y-B

Regression:

(ii) Regression equation of Y on X

$$Y - \overline{Y} = b_{yx}(X - \overline{X})$$

(i) Regression equation of X on Y

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

(ii) Regression Equation of Y on X

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

(i) Regression equation of X on Y

$$X - \overline{X} = b_{xy}(Y - \overline{Y})$$

Regression coefficient of X on Y

$$b_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$$

(ii) Regression coefficient *of* Y on X

$$b_{yx} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

1. Find the regression lines

Solution:

X	**Y**	**dx=X-12**	**dy=Y-43**	**dx^2**	**dy^2**	**dxdy**
10	40	-2	-3	4	9	6
12	38	0	-5	0	25	0
13	43	1	0	1	0	0
12	45	0	2	0	4	0
16	37	4	-6	16	36	-24
15	43	3	0	9	0	0
78	**246**	**6**	**-12**	**30**	**74**	**-18**

$$\bar{X} = \frac{\sum X}{N}$$

=78/6=13

$$\bar{Y} = \frac{\sum Y}{N} \textbf{=246/6=41}$$

$$b_{xy} = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dy^2 - \left(\sum dy\right)^2}$$

$$b_{xy} = \frac{6(-18) - (6)(-12)}{6(74) - (-12)^2}$$

b_{xy}=-0.12

$$b_{yx} = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dx^2 - \left(\sum dx\right)^2}$$

$$b_{xy} = \frac{6(-18) - (6)(-12)}{6(30) - (6)^2}$$

byx=-0.25

the regression equation of X on Y is

$(X-\overline{X})=b_{xy}(Y-\overline{Y})$

X-13=-0.12(Y-41)

X-13=-0.12Y+4.92

X=-0.12Y+4.92+13

X=-0.12Y+17.92

The regression equation of Y on X is

$(Y-\overline{Y})=b_{yx}(X-\overline{X})$

Y=-0.25X+44.25

Estimate the demand when price is 20

X=20,

Y=-0.25(20)+44.25

Y=39.25

2. **Find the mean and coefficient of correlation from the following regressionequations.**

2Y-X=50 ----- (1)

3Y-2X=10 -----(2)

2X(1) -2X+4Y=100

(2) -2X+3Y=10

(+) (-) (-)

(-)----------

Y=90

i.e., $\overline{Y}=90$

sub y=90 in (1)

2(90)-X=50

180-x=50

180-50=X

X=130

(i.e) $\overline{X} = 130$

To find the coefficient of correlation (r):

Let us assume that the regression equation of X on y is

2Y-X=50

X=2Y-50

$b_{xy}=2$

Let us assume that the regression equation of Y on X is

3Y-2X=10

3Y=2X+10

Y=(2/3)Y+(10/3)

$b_{yx}=2/3$

$r = \pm\sqrt{b_{xy} \times b_{yx}}$ = $r = \pm\sqrt{2 \times \frac{2}{3}} = 1.155$

So our assumption is wrong

Let us assume that the regression equation of Y on X is

2Y-X=50

2Y=X+50

Y=(1/2)X+25

$b_{yx}=1/2$

Let us assume that the regression equation of X on Y is

3Y-2X=10

2X=3Y-10

X=(3/2)Y-5

$b_{yx}=3/2$

$$r = \pm\sqrt{b_{xy} \times b_{yx}} \;=\; r = \pm\sqrt{\frac{1}{2} \times \frac{3}{2}} = 0.8660$$

The correlation coefficient r=0.8660

Chi – Square Test

Probability density function of χ^2 - distribution:

The probability density function of χ^2 - distribution is

$f(\chi^2) = \frac{1}{2^{\frac{v}{2}}\sqrt{\frac{v}{2}}} . (\chi^2)^{\frac{v}{2}-1} . e^{\frac{-\chi^2}{2}}$, $0 < \chi^2 < \infty$ where v is the degrees of freedom.

Important Properties of χ^2 - distribution:

(i). As degree of freedom increases v, the curve becomes more and more symmetrical. As v decreases, the curve is skewed more and more to the right.

(ii). The mean and variance of χ^2 - distribution is v and 2v respectively.

Uses:

i. χ^2 -distribution is used to test the goodness fit i.e It is used to judge whether the sample is from the hypothetical population.

ii. It is used to test the independence of attributes. i.e If a population is known to have two attributes then χ^2 -distribution is used to test whether the attributes are associated or independent based on the samples drawn from the population.

Definition: χ^2 - Test

Karl Pearson developed a test for testing the significance of discrepancy between experimental values and the theoretical values obtained under some theory or hypothesis.This test is known as χ^2 test of goodness of fit. Let o_1, o_2, ……..o_n be the observed frequencies and e_1, e_2,…….e_n be the corresponding expected frequencies such that $\sum_{i=1}^{n} o_i = N = \sum_{i=1}^{n} e_i$

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

It is a χ^2 variable with n – 1 degrees of freedom.

Conditions for the validity of χ^2 - Test:

i. The number of observations N in the sample must be reasonably large, say ≥ 50.
ii. Individual frequencies must be too small.
iii. The number of classes n must be neither too small nor too large ie $4 \leq n \leq 16$.

Problems

1. A certain drug is claimed to be effective in curing colds. In an experiment on 164 people with cold, half of them were given drug and half of them given sugar pills. The patients reaction to the treatment are recorded in the following table

	Helped	Harmed	No effect	Total
Drug	52	10	20	82
Sugar pills	44	12	26	82
Total	96	22	46	164

On the basis of this data can it be concluded that there is a significant difference in the effect of the drug and sugar pills?

Soln :

H_0 = There is no difference between the effect of the drug and sugar pills.

O	E	(O-E)	$\frac{(O-E)^2}{E}$
52	48	4	0.333
10	11	-1	0.091
20	23	-3	0.391
44	48	-4	0.333
12	11	1	0.091
26	23	3	0.391
			1.630

Degree of freedom = (r-1)(c-1) = 2
Table value = 5.991
Calculated value < Table value.
The null hypothesis is accepted.
There is no significant difference between effect of drug and sugar pills.

2. The number of automobile accidents per week in a certain community was as follows: 12 8 20 2 14 10 15 6 9 4. Are these frequencies in

agreement with the belief that accident conditions were the same during this 10 week period?

Soln :

H_0 : The given frequencies are consistent with the belief that accident conditions were the same during the 10 week period.

O	E	$\frac{(O-E)^2}{E}$
12	10	0.4
8	10	0.4
20	10	10
2	10	6.4
14	10	16
10	10	0
15	10	25
6	10	16
9	10	1
4	10	36
		26.6

Degree of freedom = (n-1) = 9

Table value = 16.9

Calculated value > Table value.

The null hypothesis is rejected.

3. The theory predicts the proportion of beans in the four groups A,B,C and D should be 9:3:3:1. In a experiment with 1600 beans the number in the four groups were 882,313,287 and 118. Does the experiment result support the theory?

Soln :

H_0 : Experimental support the theory.

Expected Frequencies

A = 900 , B = 300 , C = 300 , D = 100.

O	E	$\frac{(O-E)^2}{E}$
882	900	0.360
313	300	0.563
287	300	0.563
118	100	3.240
		4.726

Degree of freedom = (n-1) = 3

Table value = 7.81

Calculated value < Table value.
The null hypothesis is accepted.

4. A set of 5 coins is tossed 3200 times and the number of heads appearing each time is noted. The results are given below:

No of heads	0	1	2	3	4	5
Frequency	80	570	1100	900	500	50

Test the hypothesis that the coins are unbiased.

Soln :

H_0 The coins are unbiased.

Under this assumption the theoretical frequencies would follow binomial law and can be obtained by $3200(p+q)^5$

Probability of occurrence of head in a single throw = $\frac{1}{2}$

Expected frequencies can be obtained $3200\ (\frac{1}{2}+\frac{1}{2})^5$

Theoretical frequencies are 100, 500, 1000, 1000, 500, 100 respectively.

O	E	$\frac{(O-E)^2}{E}$
80	100	4
570	500	9.8
1100	1000	10
900	1000	10
500	500	0
50	100	25
		58.80

Degree of freedom = (6-1) = 5
Table value = 11.07
Calculated value > Table value.
The null hypothesis is rejected.

5. The following mistakes per page observed in a book

No of mistakes per page	No of pages
0	211
1	90
2	19
3	5
4	0
Total	325

Fit a Poisson distribution and test the goodness of fit.

Soln :

H_0 : Poisson distribution has given a good fit.

Poisson frequencies 209.43 , 92.15 , 20.27, 2.97, 0.33.

O	E	$\frac{(O-E)^2}{E}$
211	209.43	0.012
90	92.15	0.050
19	20.27	0.008
5	2.97	
0	0.33	
		0.07

Degree of freedom = (3-1) = 2

Table value = 3.84

Calculated value < Table value.

The null hypothesis is accepted.

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

	Skilled	Unskilled
Males	40	20
Females	10	30

Soln :

H_0 : Nature of work is independent of the sex of worksre.

Expected frequencies

	Skilled	Unskilled
Males	30	30
Females	20	20

O	E	$\frac{(O-E)^2}{E}$
40	30	3.333
10	20	5
20	30	3.333
30	20	5
	Total	16.666

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

Table value = 3.84

Calculated value > Table value.

The null hypothesis is rejected.

7. From the adult male population of four large cities, random sample of sizes given below are taken and the number of married and single men recorded. Do the data indicate any significant variation among the cities in the tendency of men to marry?

City	A	B	C	D	Total
Married	137	164	152	147	600
Single	32	57	56	35	180
Total	169	221	208	182	780

Soln :

H_0 : There is no significant difference in the tendency for marriage in the 4 towns.

Expected values

City	A	B	C	D	Total
Married	130	170	160	140	600
Single	39	51	48	42	180
Total	169	221	208	182	780

O	E	$\frac{(O-E)^2}{E}$
137	130	0.377
164	170	0.212
152	160	0.400
147	140	0.350
32	39	1.256
57	51	0.706
56	48	1.333
35	42	1.167
	Total	5.801

Degree of freedom = (r-1) (c-1) = 3

Table value = 7.815

Calculated value < Table value.

The null hypothesis is accepted.

There is no significant difference in the tendency for marriage in the 4 towns.

Unit IV

Inference Concerning Means & Variances

TEST OF HYPOTHESIS

Basic definitions

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

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

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

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

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

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

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

Alternative Hypothesis: Any hypothesis which is complementary to the null hypothesis is called an Alternative Hypothesis and is denoted by H1.

Errors in Sampling:

Type I and Type II errors.

Type I error : Rejection of H_0 when it is true.

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

Two types of errors occurs in practice when we decide to accept or reject a lot after examining a sample from it. They are Type 1 error occurs while rejecting Ho when it is true. Type 2 error occurs while accepting Ho when it is wrong.

Critical region: A region corresponding to a statistic t in the sample space S which lead to the rejection of Ho is called Critical region or Rejection region. Those regions which lead to the acceptance of Ho are called Acceptance Region.

Level of Significance : The probability α that a random value of the statistic "t" belongs to the critical region is known as the level of significance. In otherwords the level of significance is the size of the type I error. The levels of significance usually employed in testing of hypothesis are 5% and 1%.

One tail and two tailed test: A test of any statistical hyposthesis where the alternate hypothesis is one tailed(right tailed/ left tailed) is called one tailed test.

For the null hypothesis H_0 if $\mu = \mu_0$ then.

$H_1 = \mu > \mu_0$ (Right tail)

$H_1 = \mu < \mu_0$ (Left tail)

$H_1 = \mu \# \mu_0$ (Two tail test)

Types of samples :

Type (i): Small sample

The number of sample is less or equal to 30 is called Small sample. ie. $n \leq 30$

(Small sample test: "Students t test, F test , Chi Square test)

Type (ii): Large sample

The number of sample is above 30 is called Large sample. ie ($n > 30$)

Write short notes on critical value.

The critical or rejection region is the region which corresponds to a predetermined level of significance α. Whenever the sample statistic falls in the critical region we reject the null hypothesis as it will be considered to be probably false. The value that separates the rejection region from the acceptance region is called the critical value.

Define level of significance explain.

The probability α that a random value of the statistic 't' belongs to the critical region is known as the level of significance. In other words level of significance is the size of type I error. The levels of significance usually employed in testing of hypothesis are 5% and 1%.

LARGE SAMPLE TEST

If the size of the sample n>30 then that sample is called large sample.

Type I: Test of significance for single Mean

Type II: Test of significance for Difference of means

Type III: Test of significance for single proportion

Type IV: Test of significance for difference of proportions

TYPE V: Test of significance for difference of standard deviations.

Symbols for populations and samples:

Population size = N

Population mean = μ

Population std.deviation = σ

Population Proportion = P

Sample size = n

Sample mean = $\bar{x}$

Sample std.deviation = s

Sample proportion = $\bar{p}$

Table value:

Critical value	Level of significance		
	1%	5%	10%
Two tailed test	Z= 2.58	Z= 1.96	Z= 1.645
Right tailed test	Z= 2.33	Z= 1.645	Z= 1.28
Left tailed test	Z= -2.33	Z= -1.645	Z= -1.645

Type I : Test of significance for single Mean

Procedure

(i) Write the given data

(ii) Write the null Hypothesis and alternative Hypothesis

(iii) Write the formula and

(iv) Substitute all the given data in the formula and calculate the statistic value

(v) Write the table value

(vi) Compare the calculated Z value and the table value

(vii) Write the conclusion.

Formula:

(i) $$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

where $\bar{x}$ is the sample mean

μ is the population mean,

σ is the population std.deviation.

n is the sample size.

(ii)

$$Z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

where $\bar{x}$ is the sample mean

μ is the population mean,

s is the sample std.deviation.

n is the sample size.

Null Hypothesis:

H_0: $\mu = \mu_0$

Alternative Hypothesis:

H_1: $\mu \neq \mu_0$

Problem 1

A random sample of 200 tins of coconut oil gave an average weight of 4.95 kgs with a standard deviation of 0.21 kg. Do we accept the hypothesis of net weight 5 kgs per tin at 1% level?

Solution:

Given

Sample size n = 200

Sample mean $\bar{x}$ = 4.95

Sample std.deviation s = 0.21

Null Hypothesis: H_0: $\mu = \mu_0$

Alternative Hypothesis: H_1: $\mu \neq \mu_0$ (two tailed test)

Statistic value:

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$Z = \frac{4.95 - 5}{0.21/\sqrt{200}}$$

$$Z = -3.36$$

$$|Z| = 3.36$$

Calculated value:

$$|Z| = 3.36$$

Table value:

The table value of Z at 1% level of significance is 2.58

Conclusion:

Cal Z > Tab Z

Reject H_0

Problem 2:

A Manufacturer of ball pens claims that a certain pen the manufactures has a mean writing life of 400 pages with a standard deviation of 20 pages. A purchasing agent selects a sample of 100 pens and puts them for test. The mean writing life for the sample was 390 pages. Should the purchasing agent reject the manufactures claim at 5% level?. Table value of z at 5% level is 1.96 for two tail test and 1.64 approximately for one tail test.

Solution

Given

Sample size n = 100

Population mean μ = 400

Population std.deviation σ = 20

Sample mean $\bar{x}$ = 390

Null Hypothesis: H_0: $\mu = \mu_0$

Alternative Hypothesis: H_1: $\mu \neq \mu_0$ (two tailed test)

Statistic value:

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

$$Z = \frac{390 - 400}{20 / \sqrt{100}}$$

$$Z = -5$$

$$|Z| = 5$$

Calculated value:

$$|Z| = 5$$

Table value:

The table value of Z at 5% level of significance is 1.96

Conclusion:

Cal Z > Tab Z

Reject H_0

Problem 3:

A sample of 900 members has a mean of 3.4 cms and SD 2.61 cms. Is the sample from a large population of mean is 3.25 cm and SD 2.61 cms. If the population is normal and its mean is unknown find the 95% confidence limits of true mean.

Solution

Given

Sample size n = 900

Population mean μ = 3.25

Population std.deviation σ = 2.61

Sample mean $\bar{x}$ = 3.4

Null Hypothesis: H_0: $\mu = \mu_0$

Alternative Hypothesis: H_1: $\mu \neq \mu_0$ (two tailed test)

Statistic value:

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

$$Z = \frac{3.4 - 3.25}{2.61 / \sqrt{900}}$$

$$Z = 1.724$$

$$|Z| = 1.724$$

Calculated value:

$$|Z| = 1.724$$

Table value:

The table value of Z at 5% level of significance is 1.96

Conclusion:

Cal Z < Tab Z

Accept H_0

Type – II Test of significance for Difference of means

Consider two different normal populations with mean μ_1 and μ_2 and std, deviation σ_1 and σ_2 respectively. Let a sample size n_1 be drawn from first population and an independent sample of size n_2 drawn from second population.

Let $\bar{x}_1$ be the mean of the first sample and $\bar{x}_2$ be the mean of the second sample.

Formula:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}}$$

Where

$\bar{x}_1$ = mean of the first sample

$\bar{x}_2$ = mean of the second sample

σ_1 = std. deviation of the first population

σ_2 = std. deviation of the second population

n_1 = first sample size

n_2 = second sample size

Note 1:

If the samples have been drawn from the two population with common std.deviation

ie. $\sigma_1 = \sigma_2 = \sigma$ (say)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Note 2:

If the common std. deviation is not know

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}}$$

where

$\bar{x}_1$ = mean of the first sample

$\bar{x}_2$ = mean of the second sample

s_1 = std. deviation of the first sample

s_2 = std. deviation of the second sample

n_1 = first sample size

n_2 = second sample size

Note 3:

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_{1:} \mu_1 \neq \mu_2$

Problem 1:

A simple sample of heights of 6400 English men has a mean of 67.85 inches and SD of 2.56 inches, while a sample of heights of 1600 Australians has a mean of 68.55 inches and a SD of 2.52 inches. Do the data indicate that Americans, on the average taller than Englishmen?

Solutions:

Given

first sample size $n_1 = 6400$ second sample size $n_2 = 1600$

mean of first sample $\bar{x}_1 = 67.85$ mean of 2nd sample $\bar{x}_2 = 68.55$

std. deviation of 1st population $\sigma_1 = 2.5$ std. deviation of 2nd population $\sigma_2 = 2.52$

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (two tailed test)

The test Statistic :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}}$$

$$Z = \frac{67.85 - 68.55}{\sqrt{\frac{2.56}{6400} + \frac{2.52}{1600}}}$$

$$Z = -10$$

Calculated value:

$$|Z| = 10$$

Table value:

Table value of Z at 5% of level of significance is 1.96

Conclusion:

Cal Z > tab Z

Reject H_0

Problem 2:

The sales manager of a large company conducted a sample survey in states A and B taking 400 samples in each case. The results were in the following table.

	State A	State B
Average sales	Rs. 2,500	Rs. 2,200
S.D.	Rs. 400	Rs. 550

Test whether the average sales in the same in the 2 states at 1 % level.

Solution:

$n_1 = 400$, $\overline{x_1} = 2500$, $s_1 = 400$

$n_2 = 400$, $\overline{x_2} = 2200$, $s_2 = 550$

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$ [two tailed test]

The test statistic

$$Z = \frac{(\overline{x1} - \overline{x2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{2500 - 2000}{\sqrt{\frac{(400)^2}{400} + \frac{(550)^2}{400}}}$$

$= 8.82$

Calculated value:

$$|Z| = 8.82$$

Table value:

Table value of Z at 1% of level of significance is 2.58

Conclusion:

Cal Z > tab Z

Reject H_0

Problem3:

A college conducts both day and night classes intended to be identical. A sample of 100 day students yields examination results as $\bar{x} = 72.4$, $\sigma = 14.8$, and a sample of 200 night students as $\bar{x} = 73.9$, $\sigma = 17.9$. Are the two means statistically equal at 10% level?

Solution:

Given

first sample size n_1=100 second sample size $n_2 = 200$

mean of first sample $\bar{x}_1$ =72.4 mean of 2nd sample $\bar{x}_2$ =73.9

std. deviation of !st population σ_1 =14. std. deviation of 2nd population σ_2 =17.9

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (two tailed test)

The test Statistic :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}}$$

$$Z = \frac{72.4 - 73.9}{\sqrt{\frac{14.8}{100} + \frac{17.9}{200}}}$$

$$Z = -0.77$$

Calculated value:

$$|Z| = 0.77$$

Table value:

Table value of Z at 10% of level of significance is 1.645

Conclusion:

Cal Z < tab Z

Accept H_0

Test of Significance(Small samples)

Test of significance based on t – distribution

Definition: t - Test

Consider a normal population with mean μ and s.d σ . Let x_1, x_2, … x_n. be a random sample of size n with mean x and standard deviation s. We know that $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

standard normal variate N(0,1).

Hence the test statistics is in small sample becomes $t = \frac{\bar{x}-\mu}{\left(s\sqrt{n/n-1}\right)\sqrt{n}} = \frac{\bar{x}-\mu}{s/\sqrt{n-1}}$

Now let us define t = $\frac{\bar{x}-\mu}{s/\sqrt{n-1}}$. This follows student"s t distribution with n-1 degrees of freedom

1.**Test for the difference between the mean of a sample and that of a population**

Under the null hypothesis H_0: $\mu = \bar{x}$.

The test statistic $t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}} \sim t_{n-1}$ which can be tested at any level of significance with n-1 degrees of freedom.

II. Test for the difference between the means of two samples

II A. If x_1 and x_2 are the means of two independent samples of sizes n_1 and n_2 from a normal population with mean μ and standard deviation σ. It found that $\frac{\overline{x_1}-\overline{x_2}}{\sigma\left(\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)} \sim N(0,1)$

$t = \frac{\overline{x_1}-\overline{x_2}}{\sqrt{\left(\frac{n_1 s_1^2+n_2 s_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ which follows a t – distribution with degrees of freedom

$$\vartheta = n_1 + n_2 - 2$$

II B. When the sample sizes are equal i.e. $n_1 = n_2 = n$. The we have n pair of values. Further we have assume that the n pair are independent then the test statistic t becomes

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\left(\frac{n(s_1^2+s_2^2)}{2n-2}\right)\left(\frac{2}{n}\right)}}$$

$\therefore t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\left(\frac{(s_1^2+s_2^2)}{n-1}\right)}}$ is a student t – variate with degrees of freedom $\vartheta = 2n - 2$

II C. Suppose that the sample size are equal and if the n pairs of values in this case are not independent.

The test statistic $t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}}$ to test whether the means of differences is significantly different from zero. In this case the degrees of freedom is n – 1.

Confidence Limits (Fiducial Limits): If σ is not known and n is small then

1. 95% confidence limits for $\mu\ is\ \left(\bar{x} - \frac{st_{0.05}}{\sqrt{n-1}}, \overline{x} + \frac{st_{0.05}}{\sqrt{n-1}}\right)$
2. 99% confidence limits for $\mu\ is\ \left(\bar{x} - \frac{st_{0.01}}{\sqrt{n-1}}, \overline{x} + \frac{st_{0.01}}{\sqrt{n-1}}\right)$

Problems:

1. A sample of 10 house owners is drawn and the following values of their incomes are obtained. Mean Rs 6000, standard deviation Rs 650. Test the hypothesis that the average income of the house owners of the town is Rs 5500.

Soln :

Sample size n = 10
Sample mean $\bar{x}$ = 6000.
Population mean $\mu = 5500$
Standard deviation σ = s = 650.
$H_0 : \mu = 5500.$
$H_1: \mu \neq 5500$ (Two Tailed test).

Level of significance = 5%

Degree of freedom = n-1 = 10-1 =9
Table value = 2.262

t = $\frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}}$
S= $\frac{n}{n-1}\sigma$ = 685.16
Therefore

t = 2.189
$|t|$ = 2.189 < 2.262.
H_0 is accepted.
The average income of the house owners is Rs 5500.

2. A machinist is expected to make engine parts with axle diameter of 1.75 cm. A random sample of 10 parts shows a mean diameter of 1.85 cm with S.D of 0.1 cm. On the basis of this sample, would you say that the work of the machinist is inferior?

Sample size n = 10
Sample mean $\bar{x}$ = 1.85.
Population mean $\mu = 1.75$
Standard deviation s = 0.1.
H_0 : $\bar{x} = \mu$.
H_1: $\bar{x} \neq \mu$.
Two Tailed test).

Level of significance = 5%

Degree of freedom = n-1 = 10-1 =9
Table value = 2.262

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

Therefore
t = 3
$|t|$ = 3 > 2.262.
H_0 is rejected.

3. Samples of two types of electric bulbs were tested for length of life and the following data were obtained.

	Size	Mean	S.D
Sample I	8	1234 hrs	36 hrs
Sample II	7	1036 hrs	40 hrs

Is the difference in the mean sufficient to warrant that type I bulbs are superior to type II bulbs?

Soln :

Sample I size n_1 = 8 mean $\bar{x}_1$ = 1234 hrs s_1 = 36 hrs
Sample II size n_2 = 7 mean $\bar{x}_2$ = 1036 hrs s_2 = 40 hrs.

$H_0 : \bar{x}_1 = \bar{x}_2$

$H_1 : \bar{x}_1 > \bar{x}_2$ (Right tailed test)

Level of significance 5 %

$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S\left(\frac{1}{n1} + \frac{1}{n2}\right)}}$ $\quad$ $S = \frac{n_1 {s_1}^2 + n_2 {s_2}^2}{n_1 + n_2 - 2}$

t = 9.39

Degree of freedom v = 13

Table value = 1.77

$|t|$ = 9.39 > 1.77.

H_0 is rejected.

Type one bulbs may be regarded superior to type II bulbs.

4. A random sample of 10 boys has the following I.Q (intelligent quotients). 70, 120, 110, 101, 88, 95, 98, 107, 100. Do these data support the assumption of a population mean of a population mean I.Q of 100?

Solution:

Given n = 10, μ = 100

Set $H_0 : \mu = 100$

Under H_0, test statistics $t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{n-1}$, where $\bar{x}$ and s can be calculated from the sample data as $\bar{x}$ = 972 / 10 = 97.2 and

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n} = \frac{1833.60}{10} = 183.36$$

$$Hence\ s = 13.54$$

$$\therefore t = \frac{97.2 - 100}{13.54/9} = \frac{-2.8 \times 3}{13.54} = -6.204$$

$$\therefore |t| = 6.2\ (nearly)$$

The table value for 9 d.f at 5% level of significance is $t_{0.05}$ = 2.26

$\therefore |t| = 6.2 < t_{0.05}$. Hence the difference is not significant at 5% level. Hence H_0 may be accepted at 5% level. Hence the data support the assumption of population mean 100.

5. It was found that a machine has produced pipes having a thickness .05 mm. to determine whether the machine is in proper working order a sample of 10 pipe is chosen for which the mean thickness is .53mm and s.d is 0.3mm .test the hypothesis that the machine is in proper working order using a level of significance of (1) .05 (2) .01

Solution :

Given μ= .50, =.53; = .03; n = 10.

Set the null hypothesis H_0 :μ=50

Under H_0, test statistics $t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}} \sim t_{n-1}$, where $\bar{x}$ and s can be calculated from the sample data

$$t = \frac{0.53 - 0.50}{0.03} \times \sqrt{9} = 3.$$

(i). The table value for v = 9 d.f at 5% level of significance is $t_{0.05} = 2.26$

i. e. $|t| = 3 > t_{0.05}$

The difference is significant at 5% level of significance.

∴ The null hypothesis is rejected at 5% level of significance.

(ii). The table value for v = 9 d.f at 1% level of significance is $t_{0.01} = 3.25$.

Hence $|t| = 3 < t_{0.01}$

The difference is significant at 1% level of significance.

∴ The null hypothesis is accepted at 1% level of significance.

6. Ten soldiers participated in a shooting competition in the first week. After intensive training they participated in the competition in the second week. Their scores before and after coaching were given as follows.

Soldiers	1	2	3	4	5	6	7	8	9	10
Score before(x)	67	24	57	55	63	54	56	68	33	43
Score after(y)	70	38	58	58	56	67	68	75	42	38

Do the data indicate that the soldier have been identified by the training ?

Solution:

Here we are connected with the same set of the soldiers in the 2 competitions and their scores which are related to each other because of the intensive training .we compute the difference in their scores $z = y - x$and calculate the mean z and the s.d z as follow

x	y	$z = y - x$	$z - z$	$z - z^2$
67	70	3	-2	4
24	38	14	9	81
57	58	1	-4	16
55	58	3	-2	4
63	56	-7	-12	144
54	67	13	8	64
56	68	12	7	49
68	75	7	2	4
33	42	9	4	16
43	38	-5	-10	100
-	-	50	-	482

Given n = 10,

Under H_0, test statistics $t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}} \sim t_{n-1}$, where $\bar{x}$ and s can be calculated from the sample data
as $\bar{x}$ = 50 / 10 = 5 and

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n} = \frac{482}{10} = 48.2$$

Set the null hypothesis H_0: $\bar{x} = 0$

$$\therefore t = \frac{15}{6.94} = 2.16(nearly)$$

The table value for v = 9d.f at 55 level of significance is $t_{.05}$ = 2.26.

$\therefore |t|$ = 2.16 <t $_{.05.}$

The difference is not significant on 5% level of significance .

Hence the null hypothesis is accepted .We can conclude that there is no significant improvement in the training .

F – Test

Definition: F – Test

Let $x_1, x_2, \ldots\ldots x_n$ and $y_1, y_2, \ldots\ldots y_m$ be independent and identically distributed samples from two populations which each has a normal distribution. The expected values for the two populations can be different, and the hypothesis to be tested is that the variances are equal.

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$ be the sample means. Let $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and $S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2$ be the sample variances. Then the test statistic

$F = \frac{S_X^2}{S_Y^2}$ has an F – distribution with n -1 and m -1 degrees of freedom.

Important properties:

i. The square of the t- variate with n degrees of freedom follows a F- distribution with 1 and n degrees of freedom.

ii. The mean of the F- distribution is $\frac{v_2}{v_2-2}$, $v_2 > 2$.

iii. The variance of the F- distribution is $\frac{2\,{v_2}^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}$, $(v_2 > 4)$

v_1 & v_2 are the degrees of freedom associated with F- distribution.

Uses:

F- distribution is used to test the equality of the variance of the populations from which two small samples have been drawn.

Assumptions of F – Test:

(i)Normality: The values in each group are normally distributed.
(ii)Homogeneity: The variance within each group should be equal for all group.

Problems:

1. In one sample of 10 observations, the sum of the squares of the deviation of the sample values from the sample mean was 120 and in the other sample of 12 observations it was 314. Test whether this difference is significant at 5 % level of significance.

   Soln :

   $n_1 = 10$ $n_2 = 12$

   $\sum(X_1 - \overline{X_1})^2 = 120$ $\sum(X_2 - \overline{X_2})^2 = 314$

   $S_1^{\,2} = \frac{\sum(X_1-\overline{X_1})^2}{n_1-1} = 13.33$

   $S_2^{\,2} = \frac{\sum(X_2-\overline{X_2})^2}{n_2-1} = 28.55$

   H_0: $\sigma_1^{\,2} = \sigma_2^{\,2}$

   H_1 : $\sigma_1^{\,2} \neq \sigma_2^{\,2}$(Two tailed test)

   Degree of freedom $v_1 = 10-1 = 9$

   $v_2 = 12-1 = 11$

   F(11,9) = 3.10

   $F = \frac{s_1^{\,2}}{s_2^{\,2}} = 2.14$

   F = 2.14 < 3.10

   H_0 is accepted.

2. Two independent samples of sizes 9 and 7 from a normal population had the following values of the variables

   Sample I : 18 13 12 15 12 14 16 14 15

   Sample II :16 19 13 16 18 13 15

   Do the estimates of the population variance differ significantly at 5% level.

   Soln :

   $n_1 = 9$ and $n_2 = 7$

   $S_1^{\,2} = \frac{n_1 s_1^{\,2}}{n_1-1} = 3.751$

   $S_2^{\,2} = \frac{n_2 s_2^{\,2}}{n_2-1} = 5.2376$

   H_0: $\sigma_1^{\,2} = \sigma_2^{\,2}$

   H_1 : $\sigma_1^{\,2} \neq \sigma_2^{\,2}$(Two tailed test)

   Degree of freedom $v_1 = 9-1 = 8$

$v_2 = 7-1 = 6$

F(6,8) = 3.58

$F = \frac{{s_2}^2}{{s_1}^2} = 1.3963$

F = 1.3963 < 3.58

H_0 is accepted.

3. In comparing the variability of family income in two areas, a survey yielded the following data,

   Sample I size $n_1 = 100$ ${s_1}^2 = 25$

   Sample II size $n_2 = 110$ ${s_2}^2 = 10.$

   Assuming that the populations are normal, test the hypothesis H_0: ${\sigma_1}^2 = {\sigma_2}^2$ and H_1 : ${\sigma_1}^2 > {\sigma_2}^2$ at 5% level of significance.

   Soln :

   Sample I size $n_1 = 100$ ${s_1}^2 = 25$

   Sample II size $n_2 = 110$ ${s_2}^2 = 10.$

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

   H_1 : ${\sigma_1}^2 > {\sigma_2}^2$(right tailed test)

   Degree of freedom $v_1 = 100-1 = 99$

   $v_2 = 110-1\ 109$

   F(99,109) = 1.38

   $F = \frac{{s_1}^2}{{s_2}^2} = 2.5$

   F = 2.5 > 1.38

   H_0 is rejected .

DESIGN OF EXPERIMENTS

Analysis of variance:

The technique of analysis of variance is referred to as ANOVA. A table showing the source of variance, the sum of squares, degrees of freedom, mean squares(variance)and the formula for the "F ratio is known as ANOVA table"

The technique of analysis if variance can be classified as

(i) One way classification(CRD)
(ii) Two way classification(RBD)
(iii) Three way classification(LSD)

One way classification:

In one way classification the data are classified on the basic of one criterion

The following steps are involved in one criterion of classification

(i) The null hypothesis is

$$H_o : \mu_1 = \mu_2 = ... = \mu_k$$

$$H_1 : \mu_1 \neq \mu_2 \neq ..... \neq \mu_k$$

(ii) Calculation of total variation

Total sum of squares $V = \sum_i \sum_j x_{ij}^{\ 2} - \frac{G^2}{N}$

Where $G = \sum_i \sum_j x_{ij}$ (Grand total)

$\frac{G^2}{N}$ =correction formula

(iii) Sum of squares between the variates

$$V_1 = \sum_i \left[\frac{T_i^2}{n_i} \right] - \frac{G^2}{N}$$ With k-1 degree of freedom

(iv) Sum of squares within samples

$$V_2 = V - V_1$$

then the ratio $\frac{\frac{V_1}{K-1}}{\frac{V_2}{N-K}}$ follows F-distribution with degrees of freedom. Choosing the ratio which is greater than one, we employ the F-test

If we calculated F < table value $F_{0.05}$, the null hypothesis is accepted.

ANOVA Table for one way classification

Source of variation	Sum of square	Degrees of freedom	Mean square	Variance ratio
Between classes	V_1	K-1	$\frac{V_1}{K-1}$	$\frac{\frac{V_1}{K-1}}{\frac{V_2}{N-K}}$ (or) $\frac{\frac{V_2}{N-K}}{\frac{V_1}{K-1}}$
Within classes	V_2	$N-K$	$\frac{V_2}{N-K}$	
	V	$N-1$		

1. To test the significance of the variation of the retail prices of a certain commodity in the four principal plates A,B,C &D, seven shops were chosen at random in each city and the prices observed were as follows (prices in paise)

A	82	79	73	69	69	63	61
B	84	82	80	79	76	68	62
C	88	84	80	68	68	66	66
D	79	77	76	74	72	68	64

Do the data indicate that the prices in the four cities are significantly different?

Solution:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

i.e., the prices of commodity in the four cities are same.

we take the origin at $x = 80$ and the calculation are done as follows.

Calculation of ANOVA (use new values)

Cities K=4	Shop(n = 7) 1	2	3	4	5	6	7	T_i	$\frac{T_i^2}{n}$	$\sum x^2$
A	2	-1	-7	-11	-11	-17	-19	-64	585.14	946
B	4	2	0	-1	-4	-12	-18	-29	120.14	505
C	8	4	0	-12	-12	-14	-14	-40	228.57	760
D	-1	-3	-4	-6	-8	-12	-16	-50	357.14	526
	$\frac{G^2}{N}=1196.03$							$G=-183$	$\frac{\sum T_i^2}{n}=1290.9$	$\sum_i\sum_j x_{ij}^2=2737$

Total sum of squares $V=\sum_i\sum_j x_{ij}^2-\frac{G^2}{N}$

$=2737-1196.03$

$V=1540.97$

Sum of squares between cities

$$V_1=\sum\frac{T_i^2}{n}-\frac{G^2}{N}$$

$= 1290.9 - 1196.03$

$$V_1=94.87$$

Sum of squares within cities

$$V_2=V-V_1=1540.97-94.87$$

$$V_2=1446.1$$

ANOVA Table:

Source of variation	Sum of square of deviation	Degrees of f	Mean square	F
Between cities	$V_1=94.87$	K-1=4-1=3	$\frac{V_1}{K-1}=\frac{94.87}{3}$ $=31.62$	$=\frac{60.25}{31.62}$ $=1.90$
Within cities	$V_2=1446.1$	N-K=28-4=24	$\frac{V_2}{N-K}=\frac{1446.1}{24}$ $=60.25$	
Total	V=1540.97	N-1=27		

Number of degrees of freedom = (N - K, K - 1) = (24,3)

Critical value:

The table value of F for (24, 3) degree of freedom at 5% Los is 8.64

Conclusion:

Since F< 8.64, H_o is accepted at 5% Los

$\therefore$ The prices of commodity in the four cities are same

2. Fill up the following Analysis of variance table

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F ratio
Treatments	-	-	117	-
Error	-	704	-	
Total	16	938		

Solution:

From the given table we have,

$V_2 = 704; V=938$

degree of freedom (total) N - 1 = 16 => N = 17

mean squares $\frac{V_1}{K-1} = 117$

We Know that $V_2 = V\text{-}V_1$

$\Rightarrow V_1 = V\text{-}V_2$

$= 938 - 704$

$\boxed{V_1 = 234}$

$$\frac{V_1}{K\text{-}1} = 117$$

$$\Rightarrow \frac{234}{K\text{-}1} = 117 \Rightarrow \frac{234}{K\text{-}1} = K\text{-}1$$

$K - 1 = 2$

degree of freedom (K-1) = 2

=>K=3

Next, N-K = 17-3 = 4

$$\frac{V_2}{N\text{-}K} = \frac{938}{14} = 50.29$$

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F ratio
Treatments Error	K-1=3-1=2 N-K=17-3=14	$V_1 = 234$ $V_2 = 704$	$\frac{V_1}{K\text{-}1} = 117$ $\frac{V_2}{N\text{-}K} = 50.29$	$\frac{117}{50.29}$ = 2.327
Total	16	V = 938		

3. The following are the number of mistakes made in 5 successive days of 4 technicians working in a photographic laboratory

Technicians I	Technicians II	Technicians III	Technicians IV
6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

Test at the 1% Los whether the difference among the 4 samples means can be attributed to chance

Solution:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

ie., There is no differences among the 4 samples mean

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

We take the origin at 12 and the calculation are done as follows

Calculation of ANOVA (NEW Values)

Technicians K = 4	Days(5)					T_i	$\frac{T_i^2}{n}$	$\sum x^2$
	1	2	3	4	5			
I	-6	2	-2	-4	-1	-11	24.2	61
II	2	-3	0	-2	2	-1	0.2	21
III	-2	0	-5	3	-1	-5	5	39
IV	-3	0	-4	-2	-1	-10	20	30
Total	$\frac{G^2}{N}=\frac{(-27)^2}{20}=36.45$					G=-27	49.4	151

Total sum of squares:

$$V=\sum_i\sum_j x_{ij}^{\ 2}-\frac{G^2}{N}$$

$$=151-36.45$$

$$V=114.55$$

Sum of squares b/w cities:

$$V_1=\sum\frac{T_i^2}{n}-\frac{G^2}{N}$$

$$=49.4-36.45$$

$$V_1=12.95$$

Sum of squares within cities:

$$V_2=V\text{-}V_1=114.55-12.95$$

$$V_2=101.6$$

Source of variation	Sum of squares of deviation	Degrees of freedom	Mean squares	F ratio
B/W Technicians	$V_1 = 12.95$	K-1=4-1=3	$\frac{V_1}{K-1} = \frac{12.95}{3} = 4.31$	$= \frac{6.35}{4.31}$
Within Technicians	$V_2 = 101.6$	N-K=20-4=16	$\frac{V_2}{N-K} = \frac{101.6}{16} = 6.35$	
Total	V=114.55	N-1=19		=1.473

Degrees of freedom ((N - K, K - 1) = (16,3)

Critical value:

The table value of 'F' for (16,3) degree of freedom at 1% Los is 5.29

Conclusion:

Since F < 5.29, H_0 accepted at 1% level

$\therefore$ There is no difference among the four sample means.

4. The following table shows the lives in hours of four batches of electric lamps.

Batches	Lives in hours							
1	1610	1610	1650	1680	1700	1720	1800	
2	1580	1640	1640	1700	1750			
3	1460	1550	1600	1620	1640	1660	1740	1820
4	1510	1520	1530	1570	1600	1680		

Perform an analysis of the variance on these data and show that a significant test does not reject their homogeneity

Solution:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

I.e., the means of the lives of the four brands are homogeneous.

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

We take the origin $x_{ij} = \frac{\text{old}x_{ij} - 1700}{10}$

Calculation of ANOVA

Brand K=4	Lives								T_i	$\frac{T_i^2}{n}$	$\sum_{ij} x_{ij}$
	1	2	3	4	5	6	7	8			
1	-9	-9	-5	-2	0	2	10	-	-13	24.143	295
2	-12	-6	-6	0	5	-	-	-	-19	72.2	241
3	-24	-15	-10	-8	-6	-4	4	12	-51	325.125	1177
4	-19	-18	-17	-13	-10	-2	-	-	-79	1040.167	1247
Total	$\frac{G^2}{N} = \frac{(-162)^2}{26} = 1009.38$								G=-162	=1461.635	2960

$$N = n_1 + n_2 + n_3 + n_4 = 7 + 5 + 8 + 6 = 26$$

Total sum of squares:

$$V = \sum_i \sum_j (x_{ij})^2 - \frac{G^2}{N}$$

$$= 2960 - 1009.38$$

$$V = 1950.62$$

Total sum of squares b/w brands:

$$V_1 = \sum \frac{T_i^2}{n} - \frac{G^2}{N}$$

$$= 1461.635 - 1009.38$$

$$V_1 = 452.255$$

Sum of squares within brands:

$$V_2 = V - V_1$$

$$= 1950.62 - 452.255$$

$$V_2 = 1498.365$$

ANOVA Table:

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
B/W Brands	$V_1 = 452.255$	K-1=4-1=3	$\frac{V_1}{K-1} = \frac{452.255}{3} = 150.75$	$= \frac{150.75}{68.11}$
Within Brands	$V_2 = 1498.365$	N-K=26-4=22	$\frac{V_2}{N-K} = \frac{1498.365}{22} = 68.11$	$= 2.21$
Total	$1950.62 = V$	N-1=25		

Degrees of freedom (3, 22) =3.05

Critical value:

The table value of 'F' for (3,22) d.f at 5% Los is 3.05

Conclusion:

Since F<3.05, H_0 is accepted at 5% level

$\therefore$ The means of the lives of the four brands are homogeneous.

ie., the lives of the four brands of lamps do not differ significantly.

Two way classification:

In two way classification the data are classified on the basis of two criterions

The following steps are involved in two criterion of classification

(i) The null hypothesis

H_{01} and H_{02} framed

We compute the estimates of variance as follows

(ii) $G = \sum_i \sum_j x_{ij}$ = Grand total of $K \times n$ Observations

(iii) S : Total sum of squares $\sum\sum x_{ij}^2 - \frac{G^2}{N}$

(iv) S_1 :Sum of squares b/w rows (class-B) $= \frac{1}{K}\sum_{j=1}^{n} R_j^{\,2} - \frac{G^2}{N}$

(v) S_2 :Sum of squares b/w (classes A) $= \frac{1}{n}\sum_{i=1}^{K} C_i^{\,2} - \frac{G^2}{N}$

S_3 : Sum of squares due to error (or) Residual sum of squares

(vi) Errors (or) Residual $S_3 = S - S_1 - S_2$

(vii) The degrees of freedoms of

$S_1 = n\text{-}1$; $S_2 = k\text{-}1$; $S_3 = (n\text{-}1)(k\text{-}1)$

S=nk-1

ANOVA Table for two way classification

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
B/W 'B' classes(rows)	S_1	n-1	$\frac{S_1}{n\text{-}1} = Q_B$	$F_1 = \frac{Q_B}{Q_{AB}}$ $d.f = [(n\text{-}1)(k\text{-}1)(n\text{-}1)]$
B/W 'A' classes(column)	S_2	k-1	$\frac{S_2}{k\text{-}1} = Q_A$	$F_2 = \frac{Q_A}{Q_{AB}}$
Residual (or) error	S_3	(n-1)(k-1)	$\frac{S_3}{(n\text{-}1)(k\text{-}1)} = Q_{AB}$	$d.f = [(k\text{-}1),(k\text{-}1)(n\text{-}1)]$
Total	S	nk-1	-	-

Advantages of R.B.D:

The chief advantages of R.B.D are as follows

(i) This design is more efficient or more accurate than CRD. This is because of reduction of experimental error.

(ii) The analysis of the design is simple and even with missing observations, it is not much complicated

(iii) It is Quite flexible, any number of treatments and any number of replication may be used

(iv) It is easily adaptable as in agricultural experiment it can be accommodated well in a rectangular, squares(or)in a field of any shape

(v) It provides a method of eliminating or reducing the long term effects.

(vi) This is the most popular design with experiments in view of its simplicity, flexibility and validity. No other has been used so frequently as the R.B.D

Disadvantages:

(i) The number of treatments is very large, than the side of the blocks will increase and this may introduce heterogeneity within blocks.

(ii) If the interactions are large, the experiments may yield misleading results.

1. The following data represent the number of units of production per day turned out by four randomly chosen operators using three milling machines

Machines.

		M_1	M_2	M_3
Operators	1	150	151	156
	2	147	159	155
	3	141	146	153
	4	154	152	159

Perform analysis of variance and test the hypothesis

(i) That the machines are not significantly different

(ii) That the operators are not significantly different at 5% level

Solution:

H_{01} : There is no significantly difference bet machine and

H_{02} : There is no significantly a difference b/w operator

We take the origin 155 and the calculations are done as follows.

Calculation of ANOVA (using new values)

Operators	Machines			Row total R_j	$\sum_j x_{ij}^2$
	M1	M2	M3		
1	-5	-4	1	-8	42
2	-8	4	0	-4	80
3	-14	-9	-2	-25	281
4	-1	-3	4	0	26
Column total C_i	-28	-12	3	-37	429
$\sum_i x_{ij}^2$	286	122	21	429	

Here N=12 ; G=-37

Correction factor $\frac{G^2}{N} = \frac{(-37)^2}{12} = 114.08$

Total sum of squares:

$$S=\sum_i \sum_j x_{ij}^{\ 2} - \frac{G^2}{N}$$

$$= 429 - 114.08$$

$$= 314.92$$

Sum of squares between operators:

$$S_1 = \sum_j \frac{R_j^{\ 2}}{n_j} - \frac{G^2}{N}$$

$$= \frac{1}{3}[(-8)^2 + (-4)^2 + (-25)^2] - 114.08$$

$$= 235 - 114.08$$

$$= 120.92$$

Sum of squares between machines:

$$S_2 = \sum_i \left(\frac{C_i^{\ 2}}{n_i} \right) - \frac{G^2}{N}$$

$$= \frac{1}{4}\left[(-28)^2 + (-12)^2 + (3)^2 \right] - 114.08$$

$$= 234.25 - 114.08$$

$$S_2 = 120.17$$

Residual sum of squares:

$$S_3 = S\text{-}S_1 - S_2$$

$$= 314.92 - 120.92 - 120.17$$

$$= 73.83$$

AVOVA Table for two way classification

Source of variation	Sum of squares	Degrees of freedom	Mean sum squares	F ratio
B/W operators	120.92	n-1=4-1=3	$Q_B = \frac{S_1}{n-1} = 40.31$	
B/W machines	120.17	k-1=3-1=2	$Q_A = \frac{S_2}{k-1} = 60.09$	$\frac{40.31}{12.305} = 1.49$ $(3,6)$
Residual	73.83	(n-1)(k-1)=6	$Q_{AB} = \frac{S_3}{(k-1)(n-1)} = 12.305$	$\frac{60.09}{12.305} = 4.88$ $(2,6)$
Total	314.92	nk-1=11		

Degrees of freedom $V_1 = 2; V_2 = 6$ (machines)

Degrees of freedom $V_1 = 3; V_2 = 6$ (operators)

Critical value:

(i) Machines
The table value of 'F' for (2,6) d.f at 5% Los is 5.14

(ii) Operators
The table value of 'F' for (3,6) d,f at 5% Los is 4.76

Conclusion:

(i) Operators
Since F< 4.76 , H_{02} is accepted at 5% level
$\therefore$ The operators are not significantly different

(ii) For Machines
Since F< 5.14, H_{01} is accepted at 5% level
$\therefore$ The machines are not significantly different

2. An experiment was designed to study then performance of four different detergents, the following "whiteness" readings were obtained with specially designed equipment for 12 loads of washing distributed over three different models of washing machines.

Machines / Detergents	1	2	3	Total
A	45	43	51	139
B	47	46	52	145
C	48	50	55	153
D	42	37	49	128
Total	182	176	207	565

Looking on the detergents as treatment and the machines as blocks, obtain the appropriate analysis of variance table and test at 0.01 level of Significance whether there are differences in the detergents (or) in the washing machines

Solution:

H_{01} : There is no significant different b/w detergent

H_{02} : There is no significant different b/w washing machine

We take the origin is 50 and the calculation are done as follows.

Calculation of ANOVA (using new values)

Detergents	Washing machines			Row total R_j	$\sum_j x_{ij}^2$
	M1	M2	M3		
A	-5	-7	1	-11	75
B	-3	-4	2	-5	29
C	-2	0	5	3	29
D	-8	-13	-1	-22	234
Column total C_i	-18	-24	7	-35	367
$\sum_i x_{ij}^2$	102	234	31	367	

Here N=12; G=-35

Correction factor $\frac{G^2}{N} = \frac{(-35)^2}{12} = 102.08$

Total sum of squares: $S = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}$

$= 367 - 102.08$

S=264.92

Sum of squares b/w detergents: $S_1 = \sum_j \frac{R_j^{\,2}}{h_j} - \frac{G^2}{N}$

$$= \frac{1}{3}\left[(-11)^2 + (-5)^2 + (3)^2 + (-22)^2\right] - 102.08$$

$$= 213 - 102.08$$

$$S_1 = 110.92$$

Sum of squares between machines

$$S_2 = \sum_i \left(\frac{C_i^{\,2}}{n_i}\right) - \frac{G^2}{N}$$

$$= \frac{1}{4}\left((-18)^2 + (-24)^2 + (7)^2\right) - 102.08$$

$$= 237.25 - 102.08$$

$$S_2 = 135.17$$

Residual sum of squares $S_3 = S - S_1 - S_2$

$$= 264.92 - 110.92 - 135.17$$

$$S_3 = 18.83$$

ANOVA table for two way classification:

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
B/W detergents	$S_1 = 110.92$	n-1=4-1=3	$Q_B = \frac{S_1}{n-1} = \frac{110.92}{3} = 36.97$	$\frac{Q_B}{Q_{AB}} = \frac{36.97}{3.14} = 11.77$
B/W machines	$S_2 = 135.17$	k-1=3-1=2	$Q_A = \frac{S_2}{k-1} = \frac{135.17}{2} = 67.59$	
Residual (or) Error	$S_3 = 18.83$	(n-1)(k-1)=6	$Q_{AB} = \frac{S_3}{(n-1)(k-1)} = \frac{18.83}{6} = 3.14$	$\frac{Q_A}{Q_{AB}} = \frac{67.59}{3.14} = 21.52$
Total	S=264.92	nk-1=11		

Degrees of freedom $V_1 = 2; V_2 = 6$ (machines)

Degrees of freedom $V_1 = 3; V_2 = 6$ (detergents)

Critical value:

(i) Detergents:

The table value of F for (3,6) degree of freedom at 1% Los is 9.78

(ii)Machines

The table value of F for (2,6) degree of freedom at 1% Los is 10.92

Conclusion:

(i) For detergents

Since F>9.78, H_{01} is rejected at 5% level

$\therefore$ The detergents are significantly different

(ii) For machines

Since F>10.92, H_{02} is rejected at 5% level

$\therefore$ The machines are significantly different

3. To study the performance of three detergents and three different water temperatures the following whiteness readings were obtained with specially designed equipment.

Water temp	Detergents A	Detergents B	Detergents C
Cold Water	57	55	67
Worm Water	49	52	68
Hot Water	54	46	58

Solution:

We set the null hypothesis

H_{01} : There is no significant different in the three varieties of detergents

H_{02} : There is no significant different in the water temperatures

We choose the origin at x=50

Water temp	Detergents			Row total R_j	$\sum_j x_{ij}^2$
	A	B	C		
Cold Water	7	5	17	29	363
Worm Water	-1	2	18	19	329
Hot Water	4	-4	8	8	96
Column total C_i	10	3	43	56	788
$\sum_i x_{ij}^2$	66	45	677	788	

Total sum of squares:

$$S=\sum_j \sum_i x_{ij}^2 - \frac{G^2}{N}$$

$$= 788 - \frac{(56)^2}{9} = 788 - 348.44$$

$$S=439.56$$

Sum of squares between detergents:

$$S_1 = \sum_i \frac{C_i^2}{n_i} - \frac{G^2}{N}$$

$$= \frac{1}{3}\left[(10)^2 + (3)^2 + (43)^2\right] - 348.44$$

$$= 652.67 - 348.44$$

$$S_1 = 304.23$$

Sum of squares b/w temperatures:

$$S_2 = \sum_j \frac{R_j^2}{n_j} - \frac{G^2}{N}$$

$$= \frac{1}{3}[1266] - 348.44$$

$$= 422 - 348.44$$

$$S_2 = 73.56$$

Error sum of squares:

$$S_3 = S\text{-}S_1 - S_2$$

$$= 439.56 - 304.23 - 73.56$$

$$S_3 = 61.77$$

ANOVA Table:

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
B/W detergents	304.23	2	$\frac{304.23}{2} = 152.11$	$\frac{152.11}{15.445} = 9.848$ $(2,4)$
B/W temperatures	73.55	2	$\frac{73.56}{2} = 36.78$	$\frac{36.78}{15.445} = 2.381$
Error	61.79	4	15.445	
Total	439.56	8		

Degrees of freedom (2,4) and (2,4)

Critical value:

The table value of F for (2,4) d.f at 5% Los is 6.94

Conclusion:

(i) For detergents:

Since F > 9.85, H_{01} is rejected at 5% Los

$\therefore$ There is a significant different between the three varieties detergents,

(iii) For water temperature

Since F<6.94, H_{02} is accepted at 5% Level

$\therefore$ There is no significant different in the water temperatures.

4. Four experiments determine the moisture content of samples of a powder, each man taking a sample from each of six consignments. These assignments are

Observer	Consignment					
	1	2	3	4	5	6
1	9	10	9	10	11	11
2	12	11	9	11	10	10
3	11	10	10	12	11	10
4	12	13	11	14	12	10

Perform an analysis if variance on these data and discuss whether there is any significant different b/w consignments (or) b/w observers.

Solution:

We formulate the hypothesis

H_{02} : There is no significant different b/w observer

H_{02} : There is no significant different b/w consignment

We take origin at x=11 and the calculations are done are as follows

Calculation ANOVA:

Observer	consignments						Rowtotal R_j	$\sum_j x_{ij}^2$
	1	2	3	4	5	6		
1	-2	-1	-2	-1	0	0	-6	10
2	1	0	-2	0	-1	-1	-3	7
3	0	-1	-1	1	0	-1	-2	4
4	1	2	0	3	1	-1	6	16
Column total C_i	0	0	-5	3	0	-3	-5	37
$\sum_j x_{ij}^2$	6	6	9	11	2	3	37	

Total sum of squares $= \sum_j \sum_i x_{ij}^2 - \frac{G^2}{N}$

$$S=37-\frac{(-5)^2}{24}=35.96$$

Sum of squares b/w observers $= \sum \frac{(R_j)^2}{n_j} - \frac{G^2}{N}$

$$S_1=\frac{1}{6}\left[(-6)^2+(-3)^2+(-2)^2+(6)^2\right]-\frac{25}{24}$$

$$S_1=13.13$$

Sum of squares b/w consignments $= \sum\left(\frac{C_i^2}{n_i}\right) - \frac{G^2}{N}$

$$S_2=\frac{1}{4}\left[(0+0+25+9+9)\right]-\frac{25}{24}$$

$$S_2=9.71$$

Error sum of squares $S_3 = S - S_1 - S_2$

$$= 35.96 - 13.13 - 9.71$$

$$S_3 = 13.12$$

Source of variation	Sum of squares	Degrees of freedom	Mean squares	'F' ratio
B/W Consignments	$S_1 = 9.71$	n-1=5	$\frac{9.71}{5} = 1.94$	$\frac{1.94}{0.87} = 2.23$ (5,15)
B/W observers	$S_2 = 13.13$	k-1=3	$\frac{13.13}{3} = 4.38$	$\frac{4.38}{0.87} = 5.03$ (3,15)
Error	$S_3 = 13.12$	(n-1)(k-1)=15	$\frac{13.12}{15} = 0.87$	
Total	$S = 35.96$	nk-1=23		

Critical value:

(i) For consignments ,
The table value of 'F' for (5, 15) d.f at 5% Los is 2.90

(ii) For observers:
The table value of F for (3, 15) d,f at Los 3.29

Conclusion:

(i) For observers

Since F>3.29, H_{01} is rejected

Hence these is a difference between observers is significant

(ii) For consignment:

Since F<2.33, H_{02} is accepted

$\therefore\ \therefore$ There is no significant different b/w the consignments

LATIN SQUARES DESIGN:

A Latin squares is a squares arrangement of m-rows and m-columns such that each symbol appearly once and only once in each row and column.

In randomized block design the randomization is done within blocks the units in each block being relatively similar in L.S.D there are two restrictions

(i) The number of rows and columns are equal
(ii) Each treatment occurs once and only once in each row and column.

This design is a three way classification model analysis of variance

The following steps are involved in Latin square design

Correction factor $= \frac{G^2}{N}$; G -> Grand total

S.S b/w rows=$S_a = \sum_{i=1}^{m} \frac{S_i^2}{m} - C.F$ (S.S means Sum of Squares)

S.S b/w Columns=$S_b = \sum_{j=1}^{m} \frac{S_j^2}{m} - \frac{G^2}{N} | C.F$

S.S b/w Varieties=$S_c = \sum_{i=1}^{m} \frac{V_i^2}{m} - C.F$

$\left.\begin{matrix} Total\ sum\ of \\ squares \end{matrix}\right\} S = \sum_j \sum_i x_{ij}^2 - C.F$

and $S_d = S - S_a - S_b - S_c$

Here S_i = sum of i^{th} row

S_j = sum of j^{th} column

V_i = sum of i^{th} variety

ANOVA Table:

Source of variation	Sum of squares	Degrees of freedom	Mean squares	'F' ratio
B/W Rows	S_a	m-1	$\frac{S_a}{m-1} = R$	$\frac{R}{E}$ $[(m-1),(m-1)(m-2)]$
B/W Columns	S_b	m-1	$\frac{S_b}{m-1} = C$	$\frac{C}{E}$ $[(m-1),(m-1)(m-2)]$
B/W varieties	S_c	m-1	$\frac{S_c}{m-1} = V$	$\frac{V}{E}$ $[(m-1),(m-1)(m-2)]$
Error	S_d	(m-1)(m-2)	$\frac{S_d}{(m-1)(m-2)} = E$	
Total	S	$m^2 - 1$		

Comparison of LSD and RBD

(i) In LSD, the number of rows and number of columns are equal and hence the number of replication is equal to the number of treatments there is no such restriction in RBD

(ii) L.S.D is suitable for the case when the number of treatments is b/w 5 and 12 where as R.B.D can be used for any number of treatments and replications

(iii) The main advantage of L.S.D is that it removes the variations b/w rows and columns from that within the rows resulting in the reduction of experiment error to a large extent

(iv) The RBD can be performed equally on rectangular of square plots but for LSD, a mose (or) less a squares field is required due to (iii) LSD is preferred over RBD

Note: A 2×2 Latin Square Design is not possible. The degree of freedom for error in a $m\times m$ Latin squares design is (m-1)(m-2)

For m=2 the degree of freedom is 'o' and hence comparisons are not possible.

Hence a 2×2 LSD is not possible.

1. The following is the LSD layout of a design when 4 varieties of seeds are being tested set up the analysis of variance table and state four conclusion

A 105	B 95	C 125	D 115
C 115	D 125	A 105	B 105
D 115	C 95	B 105	A 115
B 95	A 135	D 95	C 115

Solution:

H: There is no significant difference

we take the origin as $u_{ij} = \frac{x_{ij} - 100}{5}$ and the calculations are done as follows

Varieties	Values				V_i
A	1	1	3	7	12
B	-1	1	-1	-1	0
C	5	3	-1	3	10
D	3	5	3	-1	10

Columns / Rows	C_1	C_2	C_3	C_4	Row total R_j	$\sum_i x_{ij}^2$
R_1	1	-1	5	3	8	36
R_2	3	5	1	1	10	36
R_3	3	-1	1	3	6	20
R_4	-1	7	-1	3	8	60
Columns total C_i	6	10	6	10	G=32	152
$\sum_j x_{ij}^2$	20	76	28	28	152	

G=32 N=16; $\sum_j \sum_i x_{ij}^2 = 152$

$$C.F = \frac{G^2}{N} = \frac{(+32)^2}{16} = 64$$

Total sum of squares $= \sum_{j}\sum_{i} x_{ij}^{\ 2} - \frac{G^2}{N}$

$$= 152 - \frac{(32)^2}{16}$$

$$= 152 - 64$$

$$S = 88$$

Sum of squares b/w rows $= \frac{1}{4}\left[8^2 + 10^2 + 6^2 + 8^2\right] - 64$

$$= 66 - 64$$

$$S_a = 2$$

Sum of squares b/w columns $= \frac{1}{4}\left[6^2 + 10^2 + 6^2 + 10^2\right] - 64$

$$S_b = 68 - 64$$

$$S_b = 4$$

Sum of squares b/w Varieties $= \frac{1}{4}\left[12^2 + 0^2 + 10^2 + 10^2\right] - 64$

$$= 86 - 64$$

$$S_c = 22$$

Error sum of squares $S_d = S - S_a - S_b - S_c$

$$= 88 - 2 - 4 - 22$$

$$S_d = 60$$

ANOVA Table:

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	'F' ratio
B/W rows	$S_a = 2$	m-1=4-1=3	$\frac{S_a}{m-1} = \frac{2}{3} = 0.67$	$\frac{0.67}{10} = 0.067$
B/W columns	$S_b = 4$	m-1=4-1=3	$\frac{S_b}{m-1} = \frac{4}{3} = 1.33$	$\frac{1.33}{10} = 0.133$
B/W varieties	$S_c = 22$	m-1=3	$\frac{S_c}{m-1} = \frac{22}{3} = 7.33$	$\frac{7.33}{10} = 0.733$
Error	$S_d = 60$	(m-1)(m-2) $=3\times 2=6$	$\frac{S_d}{(m-1)(m-2)} = 10$	-
Total	$S = 88$	$m^2 - 1 = 15$	-	-

Number of degrees of freedom $V_1 = 3\,; V_2 = 6$

Critical value:

The table value of F for (3, 6) d.f at 5% Los is 4.76

Conclusion:

Since F<4.76, for all the case.

$\therefore$ There is no significant difference for the varieties

2. Analyse the variance in the following Latin squares of fields (in keys) of paddy where A,B,C,D denote the difference methods of calculation

D122	A121	C123	B122
B124	C123	A122	D125
A120	B119	D120	C121
C122	D123	B121	A122

Examine whether the different methods of cultivation have given significantly different fields.

Solution:

Re arrange the table in order

A121	A122	A120	A122
B122	B124	B119	B121
C123	C123	C121	C122
D122	D125	D120	D123

We take the origin 122 and the table is

Letter	Values				V_i total
A	-1	0	-2	0	-3
B	0	2	-3	-1	-2
C	1	1	-1	0	1
D	0	3	-2	1	2

Calculation of LSD:

Columns / Rows	1	2	3	4	Row total	$\sum_j x_{ij}^2$
1	0	-1	1	0	0	2
2	2	1	0	3	6	14
3	-2	-3	-2	-1	-8	18
4	0	1	-1	0	0	2
Columns total	0	-2	-2	2	-2	36
$\sum_i x_{ij}^2$	8	12	6	10	36	

Here N=16; G=-2

Correction factor = $\frac{G^2}{N} = \frac{4}{16} = 0.25$

Total sum of squares $S = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}$

$= 36 - 0.25$

$S = 35.75$

Sum of squares b/w rows $S_a = \sum_{i=1}^{m} \frac{S_i^2}{m} - \frac{G^2}{N}$

$= \frac{1}{4}\left[(6)^2 + (-8)^2\right] - 0.25$

$= 25 - 0.25$

$S_a = 24.75$

Sum of squares b/w columns $S_b = \sum_{j=1}^{m} \frac{S_j^2}{m} - \frac{G^2}{N}$

$$= \frac{1}{4}\left[(0)^2 + (-2)^2 + (-2)^2 + (2)^2\right] - 0.25$$

$$S_b = 2.75$$

Sum of squares b/w varieties $S_c = \sum_{i=1}^{m} \frac{V_i^2}{m} - \frac{G^2}{N}$

$$= \frac{1}{4}\left[(-3)^2 + (-2)^2 + (1)^2 + (2)^2\right] - 0.25$$

$$= 4.5 - 0.25$$

$$S_c = 4.25$$

Error (or) Residual $S_d = S\text{-}S_a - S_b - S_c$

$$= 35.75 - 24.75 - 2.75 - 4.25$$

$$S_d = 4$$

LSD Table:

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	'F' ratio
B/W rows	$S_a = 24.75$	m-1=3	$\frac{S_a}{m\text{-}1} = \frac{24.75}{3} = 8.25$	$\frac{8.25}{0.67} = 12.31$
B/W columns	$S_b = 2.75$	3	$\frac{S_b}{m\text{-}1} = \frac{2.75}{3} = 0.92$	$\frac{0.92}{0.67} = 1.37$
B/W varieties	$S_c = 4.25$	3	$\frac{S_c}{m\text{-}1} = \frac{4.25}{3} = 1.42$	$\frac{1.42}{0.67} = 2.12$
Error (or) Residual	$S_d = 4.0$	6=(m-1)(m-2)	$\frac{S_d}{(m\text{-}1)(m\text{-}2)} = 0.67$	
Total	$S = 35.75$	$m^2 - 1 = 8$		

Critical value:

The value of 'F' for (3,6) d.f at 5% Los is 4.76

Conclusion:

Since F<4.76, we accept the null hypothesis

$\therefore$ The difference between the methods of cultivation is not significant.

3. The following data resulted from an experiment to compare three burners A,B, and C,A Latin squares design was used as the tests were made on 3 engines and were spread over 3 days.

	Engine 1	Engine 2	Engine 3
Day 1	A 16	B 17	C 20
Day 2	B16	C 21	A 15
Day 3	C15	A 12	B 13

Test the hypothesis that there is no diff between the burners

Solution:

We take the origin x=15 and the calculation are done as follows

Re arrangement of given table is

A 16	B 17	C 20
A 15	B 16	C 21
A 12	B 13	C 15

Varieties	Values			V_i
A	1	0	-3	-2
B	2	1	-2	1
C	5	6	0	11

Calculation of LSD

Columns/ Rows	C_1	C_2	C_3	Row total	$\sum_j x_{ij}^2$
R_1	1	2	5	8	30
R_2	1	6	0	7	37
R_3	0	-3	-2	-5	13
Column total	2	5	3	10	80
$\sum_i x_{ij}^2$	2	49	29	80	

Here N=9; G=10

Correction Factor $= \frac{G^2}{N} = \frac{(10)^2}{9} = 11.11$

Total sum of squares $S = \sum_j \sum_i x_{ij}^2 - C.F$

$= 80 - 11.11$

$S = 68.89$

Sum of squares b/w Rows $S_a = \sum_{i=1}^{m} \frac{S_i^2}{m} - C.F$

$= \frac{1}{3}[8^2 + 7^2 + (-5)^2] - 11.11$

$= 46 - 11.11$

$S_a = 34.89$

Sum of squares b/w columns $S_b = \sum_{j=1}^{m} \frac{S_j^2}{m} - C.F$

$= \frac{1}{3}[(2)^2 + (5)^2 + (3)^2] - 11.11$

$= 1.56$

Sum of squares b/w varieties $S_c = \sum_{i=1}^{m} \frac{V_i^2}{m} - C.F$

$= \frac{1}{3}[(-2)^2 + 1^2 + 11^2] - 11.11$

$S_c = 30.89$

Error (or) Residual $S_d = S - S_a - S_b - S_c$

$= 68.89 - 34.89 - 1.56 - 30.89$

$S_d = 1.55$

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	'F' ratio
B/W rows	$S_a = 34.89$	m-1=2	$\frac{S_a}{m-1} = \frac{34.89}{2} = 17.445$	$\frac{17.445}{0.775} = 22.5$
B/W columns	$S_b = 1.56$	m-1=2	$\frac{S_b}{m-1} = \frac{1.56}{2} = 0.78$	$\frac{0.78}{0.775} = 1.01$
B/W varieties	$S_c = 30.89$	m-1=2	$\frac{S_c}{m-1} = \frac{30.89}{2} = 15.445$	$\frac{15.445}{0.775} = 19.93$
Error (or) Residual	$S_d = 1.55$	(m-1)(m-2)	$S_d(m-1)(m-2) = \frac{1.55}{2} = 0.775$	
Total	$S = 68.89$	$m^2 - 1 = 8$		

Critical value:

The value of 'F' for (2,8) d.f at 5% Los is 4.46

Conclusion:

Since F> the table value for the burners

$\therefore$ There is a significant difference between the burners

and also F> tabulated F for columns the difference b/w the engines is not significant.

Homework:

1. Analyse the variance in the following LS:

B 20	C 17	D 25	A 34
A 23	D 21	C 15	B 24
D 24	A 26	B 21	C 19
C 26	B 23	A 27	D 22

\

2. Analyse the variance in the following LS:

A 8	C 18	B 9
C 9	B 18	A 16
B 11	A 10	C 20

Factorial Experiments

Definition 1:

A factorial experiment in which each of m factors at 'S' is called a symmetrical factorial experiment and is often known as S^m factorial design

Definition 2:

2^m - Factorial experiments means a symmetrical factorial experiments where each of the m-factors is at two levels

2^2 -a factorial experiment means a symmetrical experiment where each of the factors is at two levels

Note:

If the numbers of level of the different factors are equal the experiments is called as a symmetrical factorial experiment.

Uses advantages of factorial experiments:

(i) Factorial designs are widely used in experiments involving several factors where it is necessary

(ii) F.D allow effects of a factor to be estimated at several levels of the others, giving conclusions that are valid over a range of experimental conditions

(iii) The F.D are more efficient than one factor at a time experiments.

(iv) In F.D individual factorial effect is estimated with precision, as whole of the experiment is devoted to it.

(v) Factorial designs from the basis of other designs of considerable practical value.

(vi) F.D are widely used in research work. These design are used to apply the results over a wide range of conditions

2^2 -Factorial experiment:

A factorial design with two factors, each at two levels is called a 2^2 factorial design

Yates's notation:

The two factors are denoted by the letters A and B the letters 'a' and 'b' denote one of the two levels of each of the corresponding factors and this will be called the second level.

The first level of A and B is generally expressed by the absence of the corresponding letter in the treatment combinations. The four treatment combinations can be enumerated as follows.

Symbols used:

a_0b_0(or)1:Factors A and B both at first level

a_1a_0(or)a:A at second level and B at first level

a_0a_1 (or) b :A at first level and B at second level

a_1a_1(or) ab : A and B both second levels.

Yates's method of computing factorial effect totals

For the calculation of various factorial effect total for 2^2-factorial experiments the following table is need

Treatment combination	Total yield from all replicates	(3)	(4)	Effect Totals
'1'	[1]	[1]+[a]	[1]+[a]+[b]+[ab]	Grand total
a	[a]	[b]+[ab]	[ab]-[b]+[a]-[1]	[A]
b	[b]	[a]-[1]	[ab]+[b]-[a]-[1]	[B]
ab	[ab]	[ab]-[b]	[ab]-[b]-[a]+[1]	[AB]

2^2-factorial experiment conducted in a CRD

Let $x_{ij} = j^{th}$ observation of i^{th} treatment combinations i=1, 2, 3, 4; j=1,2….(say)

i.e., $x_1 = [1]; x_2 = [a]; x_3 = [b]; x_4 = [ab]$

Where

x_i=total of i^{th}treatment combination .

G=$\sum_i \sum_f x_{ij}$ grand total

n=4r=Total number of observations

$$TSS=\sum_j \sum_i x_{ij}^{\ 2} - \frac{G^2}{4r}$$

1. The follwoing table gives the plan and yields of a 2^2 – factorial experiment conducted in CRD

Analyse the design and give your comments

(1) 20	a 28	a 24	b 10
ab 23	b 11	ab 22	(1) 17
a 24	b 15	ab 21	(1) 19

Solution:

Arrange the observation as in one-way classification, we proceed as follows

Treatment Combination				Total
(1)	20	17	19	56
a	28	24	24	76
b	10	11	15	36
ab	23	22	21	66
Total			G=	234

Correction Formula = $\frac{G^2}{2^2 \times r} = \frac{234^2}{4 \times 3} = 4563$

$$\sum_j \sum_i x_{ij}^{\ 2} = 20^2 + 17^2 + 19^2 + 28^2 + 24^2 + 24^2 + 10^2 + 11^2 + 15^2 + 23^2 + 22^2 + 21^2$$

$$\sum_j \sum_i x_{ij}^{\ 2} = 4886$$

$$TSS = \sum_j \sum_i x_{ij}^{\ 2} - \frac{G^2}{4r} = 4886 - 4563 = 323$$

The values of SSA, SSB and SSAB are obtained by yate's method

Treatment combination	Total (2)	(3)	(4)	Divisor (5)	Sum of squares (6)
1	[1]	[1]+[a]	[1]+[a]+[b]+[ab]=[M]	-	-
a	[a]	[b]+[ab]	[ab]-[b]+[a]-[1]=[A]	4r	$[A]^2/4r$=SSA
b	[b]	[a]-[1]	[ab]+[b]-[a]-[1]=[B]	4r	$[B]^2/4r$=SSB
ab	[ab]	[ab]-[b]	[ab]-[b]-[a]+[1]=[AB]	4r	$[AB]^2/4r$=SSAB

SSE = TSS-(SSA+SSB+SSAB)

The analysis of variance table for 2^2 factorial design conducted in CRD

Source of variation	d.f	S.S	M.S.S	F
A	1	SSA	MSSA	$\frac{MSSA}{MSSE}$
B	1	SSB	MSSB	$\frac{MSSB}{MSSE}$
AB	1	SSAB	MSSAB	$\frac{MSSAB}{MSSE}$
Error	3(r-1)	SSE	MSSE	-
Total	4r-1	TSS	-	-

To obtain the sum of squares SSA, SSB, SSAB use yate's method:

Treatment/ combination	Total response	(3)	(4)	Divisor (5)	S.S (6)
(1)	56	56+76=132	132+102=234	4r=12	Grand total
a	76	36+66=102	20+30=50	12	$\frac{50^2}{12}=208.33$
b	36	76-56=20	102-132=-30	12	$\frac{(-30)^2}{12}=75$
ab	66	66-36=30	30-20=10	12	$\frac{(10)^2}{12}=8.33$
				Total	291.66

SSE=TSS-(SSA+SSB+SSAB)

=323-291.66

SSE=31.34

Analysis of variance table:

Source of variation	d.f	S.S	M.S.S	F	$F_{0.01}(1,6)$
A	1	208.33	208.33	53.15	13.75
B	1	75	75	19.13	
AB	1	8.33	8.33	2.09	
Error	3(r-1)=6	31.34	3.92		
Total	4r-1=11	323			

Critical value:

The table value of for (1,6) d.f at 1% Los is 13.75

Conclusion:

Since F> tabulated value of 'F' for the main effect A and B, we conclude that the main effects A and B both are significantly different at 1% Los