DIRECTORATE OF DISTANCE EDUCATION

UNIVERSITY OF JAMMU

JAMMU

SELF LEARNING MATERIAL

B.A. SEM- II

SUBJECT : STATISTICS	UNIT : I-V
	LESSON NO.: 1-22

Dr. Anuradha Goswami
COURSE CO-ORDINATOR

Editing & Proof Reading by
RAKSHA MATOO

http:/www.distanceeducationju.in

Printed and Published on behalf of the Directorate of Distance Education, University of Jammu, Jammu by the Director, DDE, University of Jammu, Jammu

PROBABILITY THEORY

UNIT-I LESSON-1

STRUCTURE

1.1 INTRODUCTION

Random variables are denoted by the last letters of the alphabet X, Y, Z, etc., with or without subscripts. For a subset B of , we usually denote by $(X \in B)$ the following event in S: $(X \in B) = (s \in S)$ X(s) $\in$ B} for simplicity. In particular, (X = x) = {s $\in$ S; X (s) = x}. The probability distribution function (or just the distribution) of an r.v. X is usually denoted by P_X and is a probability function defined on subsets of as follows: P_X (B) = P(X $\in$ B). An r.v. X is said to be of the discrete type (or just discrete) if there are countable (that is, finitely many or denumerably infinite) many points in, $x_1, x_2, \ldots$, such that $P_X(\{x_j\}) > 0$,K J ≥ 1, and $\Sigma_j P_X(\{x_j\})(= \Sigma_j P(X = x_j)) = 1$

In this lesson we have discussed discrete probability distributions viz Uniform and Bernoulli distribution.

1.2. OBJECTIVES

1. To introduce discrete probability distributions

2. To introduce uniform distribution

3. To introduce Bernoulli variate

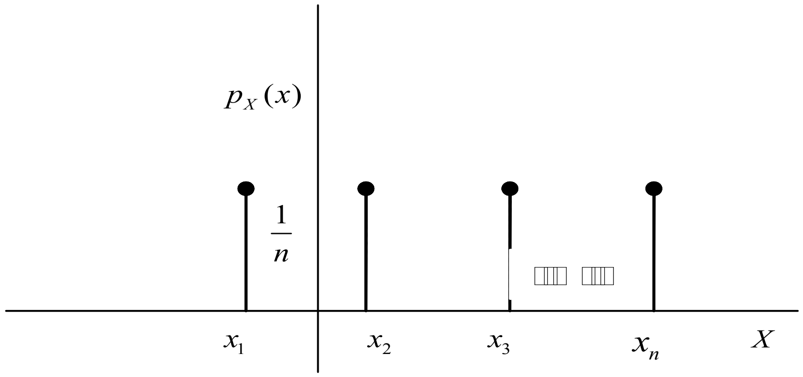
1.3 DISCRETE UNIFORM (DISTRIBUTION)

Consider a discrete random variable X with $S_X = \{x_1, x_2, ...x_n\}$. X is said to be a uniform random variable if it assumes each of the values $x_1, x_2, ...x_n$ with equal probability. The probability mass function of the uniform random variable X is given by

$$p_X(x_i) = \frac{1}{n}, \qquad i = 1, 2, ...n$$

Its CDF is

$$F_X(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$



1.3.1: Mean and Variance of the Discrete {dis (n)}

$$\mu_{\mathrm{X}} = EX = \sum_{i=0}^{n} x_i p_X(x_i)$$

$$= \frac{1}{n} \sum_{i=0}^{n} x_i$$

$$EX^2 == \sum_{i=0}^{n} x_i^2 p_X(x_i)$$

$$= \frac{1}{n} \sum_{i=0}^{n} x_i^2$$

$$\therefore \sigma_{\mathrm{X}}^2 = EX^2 - \mu_X^2$$

$$= \frac{1}{n} \sum_{i=0}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=0}^{n} x_i \right)^2$$

Example 1.3.1: Suppose X is the random variable representing the outcome of a single roll of a fair dice. Then X can assume any of the 6 values in the set $\{1, 2, 3, 4, 5, 6\}$ with the probability mass function

$$p_X(x) = \frac{1}{6} \qquad x = 1, 2, 3, 4, 5, 6$$

Example 1.3.2: Refere Example 10.2.1 above for statement

Let X = die # result. The uniform PDF is:

$$f(x) = \begin{cases} \frac{1}{6} & \text{for } x = 1,2,3,4,5,6 \\ 0 & \text{otherwise} \end{cases}$$

Notice that x_1=1, x_2=2, …, x_6=6 and $\sum_{x=1}^{6} f(x) = 1$.

Finding the mean and variance produces:

$$\mu = E(X) = \frac{\sum_{i=1}^{k} x_i}{k} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5 \qquad \text{and}$$

$$\sigma^2 = \mathrm{Var}(X) = \frac{\sum_{i=1}^{k}(x_i - \mu)^2}{k} = \frac{(1-3.5)^2}{6} + \cdots + \frac{(6-3.5)^2}{6} = 2.9167$$

1.4 BERNOULLI RANDOM VARIABLE

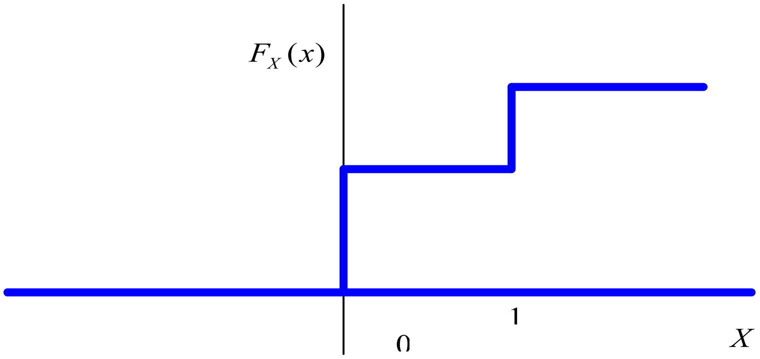
Suppose X is a random variable that takes two values 0 and 1, with probability mass functions

$p_X(1) = P\{X = 1\} = p$

and $p_X(0) = 1 - p,\qquad 0 \le p \le 1$

Such a random variable X is called a **Bernoulli random variable,** because it describes the outcomes of a **Bernoulli trial**.

The typical CDF of the Bernoulli RV X is as shown in Fig.



Remark We can define the pdf of X with the help of delta function. Thus

$f_X(x) = (1-p)\delta(x) + p\delta(x)$

Example 1.4.1: Consider the experiment of tossing a *biased* coin. Suppose $P\{H\} = p$ and $P\{T\} = 1 - p$.

If we define the random variable $X(H) = 1$ and $X(T) = 0$ then X is a Bernoulli random variable.

1.4.1. Mean and variance of the Bernoulli random variable

$$\mu_X = EX = \sum_{k=0}^{1} k p_X(k) = 1 \times p + 0 \times (1-p) = p$$

$$EX^2 = \sum_{k=0}^{1} k^2 p_X(k) = 1 \times p + 0 \times (1-p) = p$$

$$\therefore \sigma_X^2 = EX^2 - \mu_X^2 = p(1-p) = \mathrm{pq}$$

Remark

- The Bernoulli RV is the simplest discrete RV. It can be used as the building block for many discrete RVs.
- For the Bernoulli RV, $EX^m = p \quad m = 1,2,3....$ Thus all the moments of the Bernoulli RV have the same value of p.

1.5. BINOMIAL RANDOM VARIABLE:

Suppose X is a discrete random variable taking values from the set $\{0,1,.......,n\}$. X is called a binomial random variable with parameters n and $0 \le p \le 1$ if

$$p_X(k) = {}^nC_k p^k (1-p)^{n-k} \qquad k = 0,1,...,n$$

where $${}^nC_k = \frac{n!}{k!(n-k)!}$$

As we have seen, the probability of k successes in n independent repetitions of the Bernoulli trial is given by the binomial law. If X is a discrete random variable representing the number of successes in this case, then X is a binomial random variable.

For example, the number of heads in 'n' independent tossing of a fair coin is a binomial random variable.

- The notation $X \sim B(n,p)$ is used to represent a binomial RV with the parameters n and p.
- $\sum_{k=0}^{n} p_X(k) = \sum_{k=0}^{n} {}^nC_k p^k (1-p)^{n-k} = [p + (1-p)]^n = 1.$
- The sum of n independent identically distributed Bernoulli random variables is a binomial random variable.
- The binomial distribution is useful when there are two types of objects - good, bad; correct, erroneous; healthy , diseased etc.

BINOMIAL CONDITIONS

1. An experiment consists of n repeated trials.
2. Each trial has two possible outcomes: success or failure.
3. The probability of a success p is constant from trial to trial.
4. Repeated trials are independent.

Example 1.5.1.: In a binary communication system, the probability of bit error is 0.01. If a block of 8 bits are transmitted, find the probability that

(a) exactly 2 bit errors will occur

(b) at least 2 bit errors will occur

(c) more than 2 bit errors will occur

(d) all the bits will bit erroneous

Suppose X is the random variable representing the number of bit errors in a block of 8 bits. Then $X \sim B(8, 0.01)$. Therefore,

(a) Probabilty that exactly 2 bit errors will occur

$$= p_X(2)$$
$$= {}^8C_2 \times 0.01^2 \times 0.99^6$$
$$= 0.0026$$

(b) Probabilty that at least 2 bit errors will occur

$$= p_X(0) + p_X(1) + p_X(2)$$
$$= 0.99^8 + {}^8C_1 \times 0.01^1 \times 0.99^7 + {}^8C_2 \times 0.01^2 \times 0.99^6$$
$$= 0.9999$$

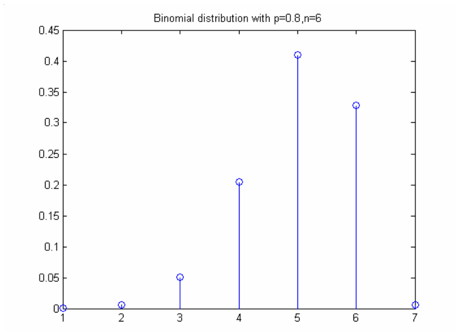
(c) Probabilty that more than 2 bit errors will occur

$$
\begin{aligned}
&= 1 - \sum_{k=0}^{2} p_X(k) \\
&= 1 - 0.9999 \\
&= 0.0001
\end{aligned}
$$

(d) Probabilty that all 8 bits will be erroneous

$$
\begin{aligned}
&= p_X(8) \\
&= 0.01^8 = 10^{-16}
\end{aligned}
$$

The probability mass function for a binomial random variable with $n = 6$ and $p = 0.8$ is shown in the figure below.



We have

1.5.1. Mean and variance of the Binomial random variable

$$
\begin{aligned}
EX &= \sum_{k=0}^{n} k p_X(k) \\
&= \sum_{k=0}^{n} k\, {}^{n}C_k\, p^k (1-p)^{n-k} \\
&= 0 \times q^n + \sum_{k=1}^{n} k\, {}^{n}C_k\, p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \frac{n-1!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-1-k_1} \\
&= np \sum_{k_1=0}^{n-1} \frac{n-1!}{k_1!(n-1-k_1)!} p^{k_1} (1-p)^{n-1-k_1} \quad \text{(Substituting } k_1 = k-1) \\
&= np(p+1-p)^{n-1} \\
&= np
\end{aligned}
$$

Thus mean of Binomial distribution is np

Similarly

$$EX^2 = \sum_{k=0}^{n} k^2 p_X(k)$$

$$= \sum_{k=0}^{n} k^2 \, {}^nC_k p^k (1-p)^{n-k}$$

$$= 0^2 \times q^n + \sum_{k=1}^{n} k \, {}^nC_k p^k (1-p)^{n-k}$$

$$= \sum_{k=1}^{n} k^2 \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

$$= \sum_{k=1}^{n} k \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k}$$

$$= np \sum_{k=1}^{n} (k-1+1) \frac{n-1!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-1-(k-1)}$$

$$= np \sum_{k=1}^{n} (k-1) \frac{n-1!}{(k-1)!(n-1-k+1)!} p^{k-1} (1-p)^{n-1-(k-1)} + np \sum_{k=1}^{n} \frac{n-1!}{(k-1)!(n-1-k+1)!} p^{k-1} (1-p)^{n-1-(k-1)}$$

$$= np \times (n-1) p + np$$

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

$$\therefore \text{ var} = n(n-1)p^2 + np - [E(x)]^2 = (n^2-n)\, p^2 + np - n^2p^2 = \cancel{n^2}p^2 - np^2 + np - \cancel{n^2}p^2$$

$$= np\,(1-p) = npq$$

Recurrance Relation for the moments of Binomial distribution

by def.

$$\mu_r = E.\{x-E(x)\}^r = \sum_{x=0}^{n} (\,x-np)^r \binom{n}{x} p^x q^{n-x}$$

diff. w.r.t p, we get

$$\frac{d\mu_r}{dp} = \sum_{x=0}^{n} \binom{n}{x} [-nr\,(x-np)^{r-1}\, p^x\, q^{n-x} + (x-np)^r \{xp^{x-1}\, q^{n-x} - (n-x)\, p^x\, q^{n-x-1}\}]$$

$$= -nr \sum_{x=0}^{n} \binom{n}{x} (x-np)^{r-1}\, p^x\, q^{n-x} + \sum_{x=0}^{n} \binom{n}{x} (x-np)^r p^x\, q^{n-x} \left\{\frac{x}{p} - \frac{n-x}{q}\right\}$$

$$= -nr \sum_{x=0}^{n} \binom{n}{x} (x-np)^{r-1}\, p\,(x) + \sum_{x=0}^{n} (x-np)^r p(x) \left\{\frac{x - np}{pq}\right\}$$

$$= -nr \sum_{x=0}^{n} \binom{n}{x} (x-np)^{r-1}\, p\,(x) + \frac{1}{pq} \sum_{x=0}^{n} (x-np)^{r+1} p(x)$$

$$\frac{d\mu_r}{dp} = -\, nru_{r-1} + \frac{1}{pq}\, \mu_{r+1}$$

$$\mu_{r+1} = pq \left[nru_{r-1} + \frac{d\mu_r}{dp} \right] \quad \textit{put r-1,2,3.....}$$

Example1.5.2.:

An accounting population contains 52 line items of which 25% are in error. A simple random sample of 6 line items is drawn. What is the probability that the sample will contain 2 items in error?

Solution:

1.

l with n = 6 and p = 0.25;

$$\text{P(2 accounts in error)} = \begin{pmatrix} 6 \\ 2 \end{pmatrix} .25^2 .75^2 = 0.2966$$

2. Sampling without Replacement

13	39
In error	Correctly state

$$\text{P(2 accounts in error)} = \frac{\begin{pmatrix} 13 \\ 2 \end{pmatrix}\begin{pmatrix} 39 \\ 4 \end{pmatrix}}{\begin{pmatrix} 52 \\ 6 \end{pmatrix}} = 0.315$$

1.6 MOMENT GENERATION FUNCTION OF BINOMIAL VARIATE

$$\phi(t) = E[e^{tX}] = \textstyle\sum_{k=0}^{n} e^{tk} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

$$= \textstyle\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) (pe^t)^k (1-p)^{n-k} = (pe^t + q)^n$$

11.7 LESSON END EXCERCISES

1. Based on past experience, a printer in a laboratory is operating 60% of the time.

Throughout a particular day, 10 visits are made and the number of times, X, that the printer is operating is observed.

(a) Write the PDF and CDF of X.

(b) What is the probability that the printer is operating:

(i) exactly 4 times?

(ii) at least 9 times?

(iii) at most 4 times?

(iv) fewer than 9 times?

2. In an accounting firm, a sample of 20 accounts is selected from a large set of accounts and subjected to examination by an auditor. The financial statements are accepted if less than two accounts are found to be in error. If 5% of all accounts are in error, what is the probability that the financial statements will be rejected?

3. Suppose you invest a fixed sum of money in each of five business ventures. Each venture has a 70% of success.

Calculate :

1. The pdf of the number of successful ventures in the five investments.
2. The expected number of successful ventures. Interpret the results.
3. The variance and the standard deviation of the number of successful ventures.

1.8. SUGGESTED READINGS

1. Goon,Gupta;Das Gupta (1991).Fundamental of Statistics
2. Gupta,S.C. and Kapoor,V.K. Fundamental of Mathematical Statistics
3. Hoel,P.G. (1971).Introductory of Mathematical Statistics
4. Hogg R.V and Craig,A.T.Introduction to Mathematical Statistics
5. Hogg ,RV and Tanis,EA(1993).Probability and Statistical Inference
6. Mood,AM,Bose DC and Graybill F A.Introduction to the Theory of Statistics
7. Rohtagi,VK.An Introduction to Probability Theory and Mathematical Statistics

****

UNIT-I LESSON-2

STRUCTURE

2.1. INTRODUCTION

The Poisson distribution is the mathematical limit to the binomial distribution and may be used to approximate binomial probabilities. The Poisson is also a distribution in its own right when solving problems involving defects per unit rather than fraction defectives. Tables showing subsets of Poisson probabilities appear in many textbooks. The tables greatly simplify the solution of many problems. The most extensive Poisson table is *Poisson's Exponential Binomial Limit* by E. C. Molina. The tables were developed in the 1920s and published in 1949.

If n is large and p is small so that n times p (np) is a positive number less than five, then the Poisson is a good approximation to the binomial. The value p and the ratio n/N should be less than 0.10.

When solving binomial problems with the Poisson formula, the terms n, x and p are the same as in the binomial formula. The task is to calculate the probability of x successes in n trials, where the probability of a single success is p. Remember that p is a fraction defective when used to approximate the binomial, andp is defects per unit when counting the number of defects instead of the number of defective units.

In some cases neither n nor p is given, but the product np may be given. If p is a fraction defective then np is the average number of defective units in the sample. If p is in terms of defects per unit then np is the average number of defects in the sample.

The Poisson probability distribution describes the number of times some event occurs during a specified interval. The interval may be time, distance, area, or volume.

The Poisson distribution is used to model many practical problems. It is used in many counting applications to count events that take place independently of one another. Thus it is used to model the *count during a particular length of time* of:

- customers arriving at a service station
- telephone calls coming to a telephone exchange packets arriving at a particular server
- particles decaying from a radioactive specimen

2.2 OBJECTIVES

1. To introduce the Poisson distribution
2. To learn the computation of mean,variance and mode of Poisson distribution
3. To introduce mgf of Poisson distribution
4. To establish recuurence relation of Poisson distribution

2.3 ASSUMPTIONS AND DEFINITION OF THE POISSON DISTRIBUTION

(1) The probability is proportional to the length of the interval.

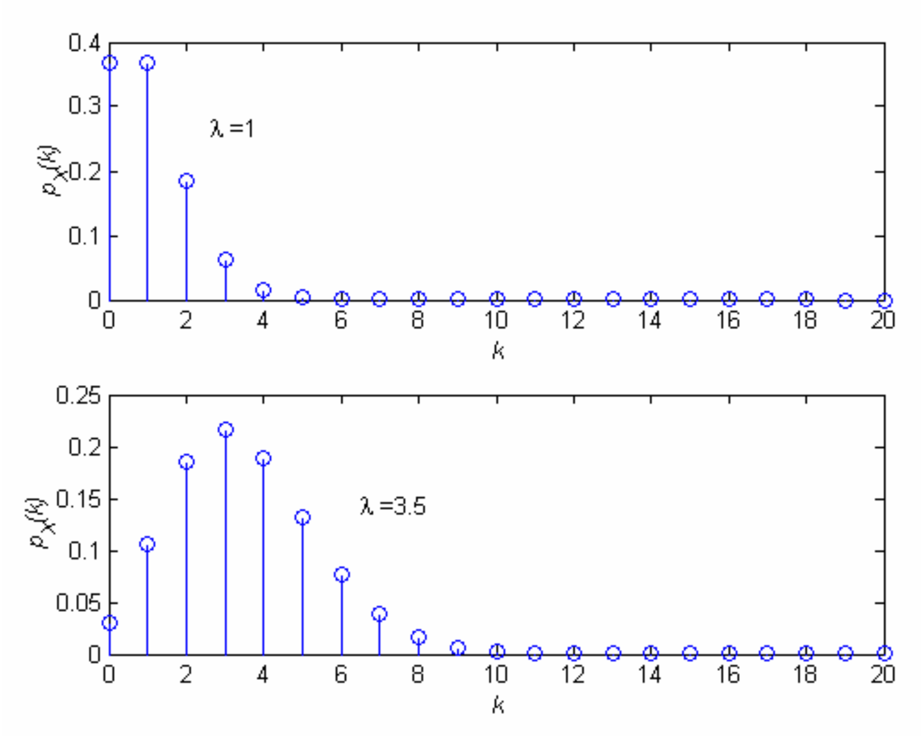
(2) The intervals are independent.

The Poisson probability distribution is characterized by the number of times an event happens during some interval or continuum.

Examples include:

- The number of misspelled words per page in a newspaper.
- The number of calls per hour received by Dyson Vacuum Cleaner Company.
- The number of vehicles sold per day at Hyatt Buick GMC in Durham, North Carolina.
- The number of goals scored in a college soccer game.

X is a Poisson distribution with the parameter λ such that $\lambda > 0$ and $p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}$, $k = 0,1,2,....$ The plot of the pmf of the Poisson RV is



Remark:

· The distribution of X is called the Poisson distribution and is denoted by P(ë). ë is called the parameter of the distribution. Often the notation X~P(ë) will be used to denote the fact that the r.v. X is distributed as P(ë).

$$p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$ satisfies to be a pmf, because

= 0 otherwise x=0,1,2,.....λ>0

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

Named after the French mathematician S.D. Poisson

2.4 MEAN AND VARIANCE OF THE POISSON DISTRIBUTION

The mean of the Poisson RV X is given by

$$\begin{aligned}
\mu_X &= \sum_{k=0}^{\infty} k p_X(k) \\
&= 0 + \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{k-1!} \\
&= \lambda e^{-\lambda} \quad \{1+\lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + ...) \\
&= \lambda e^{-\lambda} \quad e^{\lambda} = \lambda
\end{aligned}$$

∴ Hence mean of poisson distribution is λ

$$
\begin{aligned}
EX^2 &= \sum_{k=0}^{\infty} k^2 p_X(k) \\
&= 0 + \sum_{k=1}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=1}^{\infty} \frac{k\lambda^k}{k-1!} \\
&= e^{-\lambda} \sum_{k=1}^{\infty} \frac{(k-1+1)\lambda^k}{k-1!} \\
&= e^{-\lambda}\left(0 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k-2!}\right) + e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{k-1!}
\end{aligned}
$$

$$
\begin{aligned}
&= e^{-\lambda}\lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{k-2!} + e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{k-1!} \\
&= e^{-\lambda}\lambda^2 e^{\lambda} + e^{-\lambda}\lambda e \lambda \\
&= \lambda^2 + \lambda \\
\therefore \sigma_X^2 &= EX^2 - \mu_X^2 = \lambda
\end{aligned}
$$

Example 2.4.1.: The number of calls received ina telephone exchange follows a Poisson distribution with an average of 10 calls per minute. What is the probability that in one-minute duration

(i) no call is received

(ii) exactly 5 callsare received

(iii) more than 3 calls are received

Let X be the random variable~~repre~~senting the number of calls received. Given $p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}$ where $\lambda = 10$. Therefore,

(i) probability that no call is received $= p_X(0) = e^{-10} =$

(ii) probability that exactly 5 calls are received $= p_X(5) = \frac{e^{-10}\times 10^5}{5!} =$

(iii) probability that more the 3 calls are received $= 1 - \sum_{k=0}^{5} p_X(k) = 1 - e^{-10}(1 + \frac{10}{1} + \frac{10^2}{2!} + \frac{10^3}{3!}) =$

2.5 MGF OF POISSON DISTRIBUTION

If X~P(λ) then its pmf is P(X=x) = $e^{-\lambda}\ \lambda^{x}$, x=0,1,2,……

X!

$$Mx(t) = \Sigma_{x0}^{\infty}\, e^{tx}\, \frac{\lambda^x e^{-\lambda}}{x!} = \Sigma_{x=0}^{\infty}\, e^{tx}\, \frac{(\lambda e^t)^x e^{-\lambda}}{x!}$$

$$=_e^{-\lambda} \left\{ 1 + \lambda e^t + \frac{(\lambda e^t)^z}{2\iota} + ... \right\}$$

$$= e^{-\lambda} e^{\lambda e^t}$$

$$= e^{\lambda(e^t - 1)}$$

2.6 RECURRENCE RELATION OF POISSON DISTRIBUTION

If X is a random variate distributed as a Poisson distribution with parameter m then its pdf at point X=r is given by

Then its mgf is given by

$$= \frac{e^{-\lambda} + \lambda^{x+1}}{x!} \qquad x = 0,1,2,3.....$$

$$= \frac{e^{-\lambda} + \lambda^{x+1}}{(x+1)!} \qquad x = 0,1,2,3.....$$

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

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

This is the required recurrence relationship of Poisson distribution.

With this formula we can compute P(1),P(2),…….;if P(0) is known to us.It means that we can compute the probabilities at higher values if probability at some lower value is known.

2.7 POISSON DISTRIBUTION AS LIMITING CASE OF THE BINOMIAL RANDOM VARIABLE

The Poisson distribution is also used to approximate the binomial distribution $B(n, p)$ when n is very large and p is small.

Consider a binomial RV $X \sim B(n, p)$ with

$n \to \infty,\ p \to 0$ so that $EX = np = \lambda$ remains constant. Then

$$p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$\begin{aligned}
p_X(k) &= {}^nC_k\, p^k (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \frac{n(n-1)(n-2).....(n-k+1)}{k!} p^k (1-p)^{n-k} \\
&= \frac{n^k (1-\frac{1}{n})(1-\frac{2}{n}).....(1-\frac{k-1}{n})}{k!} p^k (1-p)^{n-k} \\
&= \frac{(1-\frac{1}{n})(1-\frac{2}{n}).....(1-\frac{k-1}{n})}{k!} (np)^k (1-p)^{n-k} \\
&= \frac{(1-\frac{1}{n})(1-\frac{2}{n}).....(1-\frac{k-1}{n})(\lambda)^k (1-\frac{\lambda}{n})^n}{k!(1-\frac{\lambda}{n})^k}
\end{aligned}$$

Note that $\lim_{n\to\infty}(1-\frac{\lambda}{n})^n = e^{-\lambda}$.

$$\therefore\ p_X(k) \square \lim_{n\to\infty} \frac{(1-\frac{1}{n})(1-\frac{2}{n}).....(1-\frac{k-1}{n})(\lambda)^k (1-\frac{\lambda}{n})^n}{k!(1-\frac{\lambda}{n})^k} = \frac{e^{-\lambda}\lambda^k}{k!}$$

Thus the Poisson approximation can be used to compute binomial probabilities for large n. It also makes the analysis of such probabilities easier.Typical examples are:

- number of bit errors in a received binary data file
- number of typographical errors in a printed page

Although the Poisson distribution is of independent interest, it also provides us with a close approximation of the binomial distribution for small k provided that p is small and d = np. This is indicated in the following table.

K	0	1	2	3	4	5
Binomial	.366	.370	.185	.0610	.0149	.0029
Poisson.	368	.368	.184	.0613	0153	.00307

Above is the comparison of Binomial and Poisson distributions with n = 100, p = 1/100 and X = np = 1

Example 2.7.1.

Suppose there is an error probability of 0.01 per word in typing. What is the probability that there will be more than 1 error in a page 120 words.

Suppose χ is the RV representing the number of errors per page of 120 words. $X \sim B(120, p)$ where $p = 0.01$. Therefore,

$\therefore \lambda = 120x0.01 = 0.12$

P(more than one errors)

$$= 1 - p_x(0) - p_x(1)$$
$$= 1 - e^{-\lambda} - \lambda e^{-\lambda}$$

2.8 POISSON DISTRIBUTION AS A LIMITING CASE OF NEGATIVE BINOMIAL DISTRIBUTION

Negative Binomial distribution tends to Poisson distribution as $P \to 0, r \to \infty$ such that r. $P = \lambda$ (finite).Proceeding to the limits, we get

Which is the probability function of the Poisson distribution with parameter ë.

$$\text{Lim}\,P(X{=}x) = \lim \binom{x+r-1}{r-1} p^r q^x$$

$$= \lim(\frac{x+r-1}{x})Q^{-r}(\frac{p}{Q})^x$$

$$\lim_{r\to\infty} \frac{(x+r-1)(x+r2)....(r+1).r}{x!}(1+P)^{-r}\left(\frac{p}{1+p}\right)^x$$

$$\lim_{r\to\infty} \Big[\frac{1}{x!}\left(1+\frac{\text{x-1}}{r}\right)\left(1+\frac{\text{x-2}}{r}\right) \text{........} \left(1+\frac{1}{r}\right) r^x (1+P)^{-r}\left(\frac{p}{1+p}\right)^x$$

$$= \frac{1}{x!}\lim_{r\to\infty}\left[(1+P)^{-r}\left(\frac{rP}{1+P}\right)^x\right]$$

$$= \frac{\lambda^x}{x!}\lim_{r\to\infty}\left[(1+\frac{\lambda}{r})\right]^{-r} \frac{\lim}{r\to\infty}\left[(1+\frac{\lambda}{r})\right]^{-x} \quad \therefore rP = \lambda$$

$$= \frac{\lambda^x}{x!}e^{-\lambda_{,1}}\ !$$

$$= \frac{e^{-\lambda}{}_{\lambda^x}}{x!}$$

2.9 SOME EXAMPLES

Example: Consider an inbound telemarketing operator who, on the average, handles five phone calls every three minutes. What is the probability that there will be no phone calls in the next three minutes (one unit of time)?

Solution: Let X = the number of phone calls in a time interval where a unit of time is three minutes. The l is 5 for the ONE unit of time (3 minutes). Thus, lt = 5*1 = 5. Then

$$\text{P(X = 0) = (0) = }\frac{5^0 \text{e}^{-5}}{0!}\text{ = 0.0067}$$

What is the probability that there will be no phone calls in the next minute?

Since there is only one minute, we have only 1/3 of a unit of time. Then $\lambda t = 5/3 = 41.67$

Let X = the number of phone calls. Then

$$f(0) = \frac{(5/3)^0 e^{-(5/3)}}{0!} = e^{-(5/3)} = 0.1889$$

What is the probability that there will be 2 or more phone calls in the next minute?

$P(X\ {}^{3}\ 2) = P(X=2) + P(X=3) + P(X=4) + P(X=5) + P(X=6) + \ldots$

$= 1 - P(X=0) - P(X=1)$

$$= 1 - \frac{(5/3)^0 e^{-(5/3)}}{0!} - \frac{(5/3)^1 e^{-(5/3)}}{1!}$$

$= 1 - (0.1889 + 0.3148)$

$= 1 - 0.5037$

$=0.4963$

Example 2.9.2 Suppose that we are investigating the safety of a dangerous intersection. Past police records indicate a mean of 5 accidents per month at this intersection. Suppose the number of accidents is distributed according to a Poisson distribution. Calculate the probability in any month of exactly 0, 1, 2, 3 or 4 accidents.

Solution : Since the number of accidents is distributed according to a Poisson distribution and the mean number of accidents per month is 5, we have the probability of happening accidents in any month $p(x) = \frac{5^x e^{-5}}{x!}$. By this formula we can calculate

$p(0) = 0.00674$, $p(1) = 0.3370$, $p(2) = 0.08425$, $p(3) = 0.14042$, $p(4) = 0.17552$.

The probability distribution of the number of accidents per month is presented in Table

Poisson probability distribution of the number of accidents per month

X- NUMBER OF ACCIDENTS	*P*(X) - PROBABILITY
0	0.006738
1	0.03369

2	0.084224
3	0.140374
4	0.175467
5	0.175467
6	0.146223
7	0.104445
8	0.065278
9	0.036266
10	0.018133
11	0.008242
12	0.003434

2.10 LESSON END EXERCISE:

1. If the random variable X has a Poisson distribution such that Pr (X = 1) = Pr (X = 2), find Pr (X = 4).

2. In a lengthy manuscript, it is discovered that only 13.5 per cent ofthe pages contain no typing errors. If we assume that the number of errors per page is a random variable with a Poisson distribution, find the percentage of pages that have exactly one error.

3. Let the number of chocolate drops in a certain type of cookie havea Poisson distribution. We want the probability that a cookie of this type contains at least two chocolate drops to be greater than 0.99. Findthe smallest value that the mean of the distribution can take.

2.11 SUMMARY:

Characteristics defining a Poisson random variable

1. The experiment consists of counting the number *x* of times a particular event occurs during a given unit of time

2. The probability that an event occurs in a given unit of time is the same for all units.

3. The number of events that occur in one unit of time is independent of the number that occur in other units.

The mean number of events in each unit will be denoted by the Greek letter λ

The probability distribution, mean and variance for a Poisson random variable *x:*

1. The probability distribution:

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

where λ = mean number of events during the given time period,

e = 2.71828...(the base of natural logarithm).

2) The mean: $\mu = \lambda$

3) The variance: $\sigma^2 = \lambda$

2.12 SUGGESTED READINGS:

1. Goon,Gupta;Das Gupta (1991).Fundamental of Statistics
2. Gupta,S.C. and Kapoor,V.K. Fundamental of Mathematical Statistics
3. Hoel,P.G. (1971).Introductory of Mathematical Statistics
4. Hogg R.V and Craig,A.T.Introduction to Mathematical Statistics
5. Hogg ,RV and Tanis,EA(1993).Probability and Statistical Inference
6. Mood,AM,Bose DC and Graybill F A.Introduction to the Theory of Statistics
7. Rohtagi,VK.An Introduction to Probability Theory and Mathematical Statistics

******

UNIT-I LESSON-3

STRUCTURE

3.1 INTRODUCTION

In this lesion we have discussed the need and situations when Negative binomial distribution(NBD) is applicable.We have defined the NBD and derived its moments in this lesion.Further we have derived the expression for mgf of NBD.As recurrence is used to derive the probabilities for higher values of random variables.We have also established the recurrence relation for NBD**.**

3.2 OBJECTIVES

1. To introduce the Negative Binomial distribution

2. To derive the expressions for mean and variance of NBD

3. To establish recurrence relation of NBD

4. To show Poisson distribution as limiting case of NBD

3.3 NEGATIVE BINOMIAL DISTRIBUTION

In its simplest form, the negative binomial distribution models the number of successes before a specified number of failures is reached in an independent series of repeated identical trials. It can also be thought of as modelling the total number of trials required before a specified number of successes, thus motivating its name as the "inverse" of the binomial distribution. Its parameters are the probability of success in a single trial, p , and the number of failures, r . A special case of the negative binomial distribution, when $r = 1$, is the geometric distribution (also known as the Pascal distribution), which models the number of successes before the first failure.

More generally, the r parameter can take on non-integer values. This form of the negative binomial has no interpretation in terms of repeated trials, but, like the Poisson distribution, it is useful in modelling count data. It is, however, more general than the Poisson, because the negative binomial has a variance that is greater than its mean, often making it suitable for count data that do not meet the assumptions of the Poisson distribution. In the limit, as the parameter r increases to infinity, the negative binomial distribution approaches the Poisson distribution.

Definition : A random variable x is said to follow a negative binomial distrubition if its probability mass funcuton is given by $p(x) = p\ (x=x) = \underset{r-1}{\underline{(x + r- 1)}}\ p^r q^x \quad x=0,1,2,3......$

Notation :

The following notation is helpful, when we talk about negative binomial probability.

* *x*: The number of trials required to produce *r* successes in a negative binomial experiment.

* *r*: The number of successes in the negative binomial experiment.

* *p*: The probability of success on an individual trial.

* *q*: The probability of failure on an individual trial. (This is equal to 1 - *P*.)

* b*(*x*; *r, P*): Negative binomial probability - the probability that an *x*-trial negative binomial experiment results in the *rth* success on the *xth* trial, when the probability of success on an individual trial is *P*.

* **$_n$C$_r$:** The number of combinations of *n* things, taken *r* at a time.

Properties of NBD:

A **negative binomial experiment** is a statistical experiment that has the following properties:

- The experiment consists of *x* repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
- The probability of success, denoted by *P*, is the same on every trial.
- The trials are independent that is, the outcome on one trial does not affect the outcome on other trials.
- The experiment continues until *r* successes are observed, where *r* is specified in advance.

The mean for the Negative Binomial Distribution is **rp**.

The variance for the Negative Binomial Distribution is **rqp.**

When the r parameter is an integer, the negative binomial pdf is

$$y = f(x/r, p) = \left(\frac{r+x-1}{x}\right) p^r q^x \mathrm{I}_{(0,1.....)^{(x)}}$$

where $q = 1 - p$. When r is non-integer, the binomial coefficient in the definition of the pdf is replaced by the equivalent expression

$$\frac{\Gamma(r+x)}{\Gamma(r)\Gamma(x+1)}$$

Consider the situation where one performs a number Bernoulli trials, each trial has a probability of success p , andtrials continue until the r thsuccesss occurs. Let X be the random variable which is the number of trials up toand including the r th success. This means that the range of X is the set $\{r, r+1, ...\}$. Then the pmf would be given by $\left(\frac{x-1}{r-1}\right)p^r q^{x-r}$.

Note that there are r successes and x-r previous failures, with the last success at a fixed rthposition. Thus the number of possible outcomes is the number of combinations of selecting $(x-1)$ objects taken $(r-1)$ at a time.

3.4 THE MOMENT GENERATING FUNCTION OF THE NEGATIVE BINOMIAL DISTRIBUTION

$$M_{X^{(t)}} = E(e^{tX}) = \sum_{x=o}^{\infty} e^{tx}\ p(x)$$

$$= \sum_{x=o}^{\infty} \frac{-r}{x}\ Q^{-r}\left(\frac{-Pe^{t}}{Q}\right) = (Q\text{-}Pe^{t})^{-r}$$

3.5. THE MEAN OF THE NEGATIVE BINOMIAL DISTRIBUTION

$Mx^{(t)} = E\ (e^{tx})$

$= (Q - Pe^{t})^{-r}$

$$\mu 1^{1} = \left(\frac{d}{dt}\ m(t)\right)_{t=0}$$

$\mu 1^{1} = [-r\ \ (-pe^{t})\ (Q\text{-}pe^{t})^{r-1}]_{t=0}$

$= rp$

3.6 THE VARIANCE OF THE NEGATIVE BINOMIAL DISTRIBUTION

The second moment of E(X2) of the negative binomial distribution with the pmf and mgf above is obtained by differentiating the mgf twice wrt and setting t to zero and the variance is computed as

$$\mu_2^1 = \left(\frac{d^2\, m(t)}{dt^2} \right)_{t=0}$$

$$=[r\, pe^t\, (Q-pe^t)^{r-1}]+(-r-1)r\, Pe^t\, (Q - pe^t)^{-r-2}\, (-pe^t)_{\;t=0}$$

$$= rp + (r+1)\, P^2$$

$$\therefore\ \mu_2 = \mu_2{}^1 - \mu_1{}^{1^2} = r\,(r+1)\, P^2 + rp - r^2p^2 - rPQ$$

Recurrence formula for N.B.D. is $f(x+1; r,p) = \frac{x+r}{x+1}\, q\, f(x;r,p)$

$$f(x; r,p) = \frac{(x+r-1)}{r-1}\, p^r q^x$$

$$f(x+1\, ; r,p) = \frac{(x+r)}{r-1}\, p^r q^{x+1}$$

$$\therefore\ \frac{f(x+1\, ;\, r,p)}{f\quad (x\, ;\, r,p)} = \frac{(x+r)!\ (r-1)\ !\ \ (x!)}{(r+1)!\ (x-1)\ !\ \ (x+r-1)}\quad p = \frac{x+r}{x+1}\, q$$

$$\rightarrow f\,\frac{(x+1\, ;\, r,p)}{} = \frac{(x+r)}{(x+1)}\, q\quad f\,\underline{x;\, r,p}$$

3.7 EXAMPLES OF NEGATIVE BINOMIAL DISTRIBUTION:

The distribution of X is called the Negative Binomial distribution. This distribution occurs in situations which have as a model the following. A Binomial experiment E, with sample space {S, F}, is repeated independently until exactly r S's appear and then it is terminated. Then the r.v. X represents the number of times beyond r that the experiment is required to be carried out, and f(x) is the probability that this number of times is equal to x. In fact, here S ={all (r + x)-sequences of S's and F 's such that the rth S is at the end of the sequence}, x = 0, 1, . . . and f(x) = P(X = x) = P[all (r + x)-sequences as above for a specified x]. The probability of one such sequence is $p^{r-1}q^xp$ by the independence assumption,we can derive its pdf.

$$P(X = r) = {}_{n-1}C_{r-1}\ p^{\,r}\,(1-p)^{n-r}$$

The above interpretation also justifies the name of the distribution. For r = 1,we get the Geometric (or Pascal) distribution, namely f(x) = pqx, x = 0, 1, 2, .

The probability distribution of a Negative Binomial random variable is called a Negative Binomial Distribution. It is also known as the Pascal distribution or Polya distribution. Suppose we flip a coin repeatedly and count the number of heads (successes). If we continue flipping the coin until it has landed 2 times on heads, we are conducting a Negative Binomial Experiment.

$P(X = r) = {}^{n-1}C_{r-1}\ p^{r}(1-p)^{n-r}$ where, n = Number of events. r = Number of successful events. p = Probability of success on a single trial. ${}^{n-1}C_{r-1} = ((n-1)! / ((n-1)-(r-1))!) / (r-1)!$ 1-p = Probability of failure.

Example 3.7.1: Find the probability that a man flipping a coin gets the fourth head on the ninth flip.

Solution:

Step 1: Here,

Number of trials n = 9 (because we flip the coin nine times).

Number of successes r = 4 (since we define Heads as a success).

Probability of success for any coin flip p = 0.5

Step 2: Find n-1 and r-1.

n-1 = 9-1 = 8

r-1 = 4-1 = 3

Step 3: To find ${}^{n-1}C_{r-1}$ Calculate ((n-1)-(r-1))!

(n-1)-(r-1) = 8-3 = 5

((n-1)-(r-1))! = 5! = 120

Step 4: Find (n-1)!

= 8! = 40320

Step 5: Find (r-1)!

$= 3! = 6$

Step 6: Find (n-1)! / ((n-1)-(r-1))!

$= 40320/120 = 336$

Step 7: To Solve ${}_{n-1}C_{r-1}$ formula is used.

$= 336/6 = 56$

Step 8: Find p^r.

$= 0.5^4 = 0.0625$

Step 9: To Find $(1-p)^{n-r}$ Calculate 1-p and n-r.

$1-p = 1-0.5 = 0.5$

$n-r = 9-4 = 5$

Step 10: Calculate $(1-p)^{n-r}$.

$= 0.5^5 = 0.03125$

Step 11: Calculate Negative Binomial Distribution.

$= 56\times0.0625\times0.03125 = 0.109375$

The probability that the coin will land on heads for the fourth time on the ninth coin flip is 0.1094

Example 3.7.2.

Bob is a high school basketball player. He is a 70% free throw shooter.That means his probability of making a free throw is 0.70. During the season, whatis the probability that Bob makes his third free throw on his fifth shot?

Solution:

This is an example of a negative binomial experiment. The probability of success (p) is 0.70, the number of trials (x) is 5, and the number of successes (r) is3.

To solve this problem, we enter these values into the negative binomial formula.

$f(5) = 4_{C_2} * 0.73 * 0.32$

$= 6 * 0.343 * 0.09 = 0.18522$

Thus, the probability that Bob will make his third successful free throw on his fifthshot is 0.18522.

Example 3.7.3

Let's reconsider the above problem from Example 1. This time, we'll ask a slightlydifferent question: What is the probability that Bob makes his first free throw onhis fifth shot?

Solution:

This is an example of a geometric distribution, which is a special case of a negative binomial distribution. Therefore, this problem can be solved using thenegative binomial formula or the geometric formula.The probability of success (*p*) is 0.70, the number of trials (*x*) is 5, and the number of successes (*r*) is 1. We enter these values into the negative binomial formula.

f(5) = 4_{C_0} * 0.71* 0.34

= 0.00567

Now, we demonstrate a solution based on the geometric formula.

g(5; 0.7) = 0.7 * 0.34

= 0.00567

3.7.1 Alternative Views of the Negative Binomial Distribution

As if statistics weren't challenging enough, the above definition is not the only definition for the negative binomial distribution. Two common alternative definitions are:

- The negative binomial random variable is *R*, the number of successes before the binomial experiment results in *k* failures. The mean of *R* is:

  ì$_{R}$ = *kP*/*Q*

- The negative binomial random variable is *K*, the number of failures before the binomial experiment results in *r* successes. The mean of *K* is:

  ì$_{K}$ = *rQ*/*P*

The moral: If someone talks about a negative binomial distribution, find out how they are defining the negative binomial random variable.

On this web site, when we refer to the negative binomial distribution, we are talking about the definition presented earlier. That is, we are defining the negative binomial random variable as *X*, the total number of trials required for the binomial experiment to produce *r* successes.

3.8 LESSON END EXERCISE

1. Find the moment generating function of Negative binomial distribution.Find its mean and variance from its mgf.

2. If a fair coin is tossed at random five independent times, find the conditional probability of five heads relative to the hypothesis that there are at least four heads.

3. Let X be an r.v. distributed as Negative Binomial with parameters rand p. By working as in the Binomial example, show that $EX = rq/p$, $\sigma^2(X) = rq/p^2$;

4. For each example state whether or not the negative binomial distribution is appropriate, briefly explain why or why not.

   1. Number of cars passing along a road until five red cars have passed.

   2. Number of cars passing along a road until five commercial vehicles have passed.

   3. Number of cars passing along a road until five have passed with at least one of the letters X, Y, Z in their number plate.

   4. Number of cars passing along a road until 5 have passed containing only the driver.

   5. Number of people who enter a shop until 5 men have entered.

   6. Number of children entering a school until 5 enter with glasses have entered.

   7. Number of tosses of a coin until 5 tails have appeared.

   8. Number of cards dealt from a pack of 52 cards until there are 5 black cards

3.9. SUGGESTED READINGS

1. Goon,Gupta;Das Gupta (1991).Fundamental of Statistics
2. Gupta,S.C. and Kapoor,V.K. Fundamental of Mathematical Statistics
3. Hoel,P.G. (1971).Introductory of Mathematical Statistics
4. Hogg R.V and Craig,A.T.Introduction to Mathematical Statistics
5. Hogg ,RV and Tanis,EA(1993).Probability and Statistical Inference
6. Mood,AM,Bose DC and Graybill F A.Introduction to the Theory of Statistics
7. Rohtagi,VK.An Introduction to Probability Theory and Mathematical Statistics

******

UNIT-I LESSON-4

STRUCTURE

4.1. INTRODUCTION

In probability theory the **Hypergeometric distribution (HGD)** is a discrete probability distribution that describes the number of successes in a sequence of *n* draws from a finite population without replacement.In this lesson we have discussed the HGD and its properties.

4.2. OBJECTIVES

1. To introduce the situations where HGD is applicable

2. To derive the moments of HGD

3. To establish association of HGD with Binomial distribution

4.3. HYPER GEOMETRIC DISTRIBUTION

A typical example of HGD is illustrated by this contingency table Contingency, table:

There is a shipment of N objects in which m are defective. The hypergeometric distribution describes the probability that in a sample of n distinctive objects drawn from the shipment exactly k objects are defective.

	drawn	**not drawn**	**Total**
defective	k	m-k	m
non-defective	n-k	N+k-n-m	N-
total	mn	N-n	N

In general, if a random variable follows the hypergeometric distribution with parameters N, m and n, then the probability of getting exactly k successes is given by

$$f(k;N,m,n)=\frac{\binom{m}{k}\binom{N-m}{n-k}}{\genfrac{}{}{0pt}{}{N}{n}}$$

$N \in 1,2,3,.......$

$m \in 0,1,........,N$

$n \in 1,2,........,N$

The probability is positive when k is between max$\{0,n + m - N\}$ and min$\{m,n\}$.

The classical application of the Hypergeometric distribution is **sampling without replacement**. Think of an urn with two types of marbles, black ones and white ones. Define drawing a white marble as a success and drawing a black marble as a failure (analogous to the binomial distribution). If the variable N describes the number of **all marbles in the urn** (see contingency table above) and m describes the number of **white marbles** (called *defective* in the example above), then N - m corresponds to the number of **black marbles**.

Defination of Hypergeometric Distribution

A discrcte random variable x is said to follow. the hyper Geometric distribution if it assumes only non-negative values & its probability mass functions is given by

$$P\ (x=K)\ h(K;\ N,\ M,\ n) = \frac{\binom{M}{K}\binom{N-M}{n-K}}{\binom{N}{n}} \quad K=0,1,2.........\ \min(n,m)$$

X is said to have a **hypergeometric distribution**

Example 4.3.1:

Draw 6 cards from a deck without replacement. What is the probability of getting two hearts?

Solution

Here M = 13 number of hearts

L = 39 number of non-hearts

N = 52 total

$$P(2\text{ hearts}) = \frac{\binom{13}{2}\binom{39}{4}}{\binom{52}{6}} = 0.315$$

Example 4.3.2: Lotto

42 balls are numbered 1 - 42. You select six numbers between 1 and 42. (The ones you write on your lotto card). 6 balls are selected at random. What is the probability that they contain:

(i) 4 of yours?

(ii) 5 of yours? Answer = 0.00004

(iii) 6 of yours? Answer = 0.00000019

Solution (i)

Total = 42; N = 42

Favourable = 6; M = 6

Non-Favourable = 36. L = N - M = 42 - 6 = 36

Sample size n = 6 and x = 4.

$$P(\text{match } 4) = \frac{\begin{pmatrix}6\\4\end{pmatrix}\begin{pmatrix}36\\2\end{pmatrix}}{\begin{pmatrix}42\\6\end{pmatrix}} = .0018$$

So odds of about 1 in 555.

4.4. SAMPLING WITHOUT REPLACEMENT

While the binomial distribution is obtained while sampling **with** replacement, the hypergeometric distribution is obtained while sampling **without** replacement.

4.4.1. Example:

An accounting population contains 52 line items of which 25% are in error. A simple random sample of 6 line items is drawn. What is the probability that the sample will contain 2 items in error?

Solution:

1. **Sampling with Replacement $\Rightarrow$ Binomial**

Binomial with n = 6 and p = 0.25;

$$\text{P(2 accounts in error)} = \begin{pmatrix}6\\2\end{pmatrix}.25^2.75^4 = 0.2966$$

2. **Sampling without Replacement**

13 In error   39 Correctly state

$$\text{P(2 accounts in error)} = \frac{\begin{pmatrix}13\\2\end{pmatrix}\begin{pmatrix}39\\4\end{pmatrix}}{\begin{pmatrix}52\\6\end{pmatrix}} = 0.315$$

4.5. RELATIONSHIP OF HYPERGEOMETRIC DISTRIBUTION WITH BINOMIAL DISTRIBUTION:BINOMIAL OR HYPERGEOMETRIC?

Approximation to Binomail distribution : hypergeometric distribution tends to binomial distribution as N→ ∝ and $\frac{M}{N} \to P$

$$n(K;N,M,n) = \binom{M}{K}\binom{N-M}{n-K} \div \binom{N}{n}$$

$$= \frac{M!}{K!(M-K)!} \;,\; \frac{(N-M)!}{(n-K)!(N-M-N+K)!} \;,\; \frac{n!(N-n)!}{N!}$$

$$= \frac{M(M-1)(M-2)\ldots\ldots(M-K+1)}{K!} \times \frac{(N-M)N-M-1)\ldots\ldots(N-M-n+K+1}{(n-K)!} \times \frac{n!}{N(N-1)(N-2)\ldots\ldots N-n+1}$$

$$= \frac{n!}{K!(n-K)!} . \frac{\frac{M}{N}\left(\frac{M}{N}-\frac{1}{N}\right)\left(\frac{M}{N}-\frac{2}{N}\right)\ldots\ldots\left(\frac{M}{N}-\frac{K-1}{N}\right) \times \left(1-\frac{M}{N}\right)\left(1-\frac{M}{N}-\frac{1}{N}\right)\ldots\ldots\left(1-\frac{M}{N}-\frac{n-K-1}{N}\right)}{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\ldots\ldots\left(1-\frac{n-1}{N}\right)}$$

Proceding to the limit as N→∝ & Putting $\frac{M}{N} = p$, we get lt $h_{n\to\propto}(K;N,M,n) = \binom{n}{k}p$

p ... P (1-p) (1-p) ...... (1-P) $\binom{n}{k} p^K (1-p)^{n-k}$ N → ∝

What is the probability of getting no more than 2 misstated accounts in a simple random sample of size 10 drawn without replacement from an accounting population that has a 10% misstatement rate.

Let X = number of incorrectly stated accounts.

P(X = 2) = P(X = 0) + P(X = 1) + P(X = 2)

Hypergeometric

For a population of size 20

$$P(X=0)=\frac{\binom{2}{0}\binom{18}{10}}{\binom{20}{10}}=.2368$$

$$P(X=1)=\frac{\binom{2}{1}\binom{18}{9}}{\binom{20}{10}}=.5263$$

$$P(X=2)=\frac{\binom{2}{2}\binom{18}{8}}{\binom{20}{10}}=.2368$$

Population Size	**20**
P(X = 0)	.2368
P(X = 1)	.5263
P(X = 2)	.2368
P(X = 2)	**.9999**

Hypergeometric (without replacement sampling)

For a population of size 200

$$P(X=0)=\frac{\binom{180}{10}}{\binom{200}{10}}=.3398$$

$$P(X=1)=\frac{\binom{20}{1}\binom{180}{9}}{\binom{200}{10}}=.3974$$

$$P(X=2)=\frac{\binom{20}{2}\binom{180}{8}}{\binom{200}{10}}=.1975$$

P(X = 2) = .9347

Population Size	**200**
P(X = 0)	.3398
P(X = 1)	.3974
P(X = 2)	.1975
P(X = 2)	**.9347**

Hypergeometric (without replacement sampling)

For a population of size 2,000

$$P(X=0)=\frac{\binom{1800}{10}}{\binom{2000}{10}}=.3476$$

$$P(X=1)=\frac{\binom{200}{1}\binom{1800}{9}}{\binom{2000}{10}}=.3881$$

$$P(X=2)=\frac{\binom{200}{2}\binom{1800}{8}}{\binom{2000}{10}}=.1939$$

P(X = 2) = .9296

Population Size	**2,000**
P(X = 0)	.3476
P(X = 1)	.3881
P(X = 2)	.1939
P(X = 2)	**.9296**

Binomial (with replacement sampling)

$$P(X=0)=0.9^{10}=.3487$$

$$P(X=1)=\binom{10}{1}\,0.1^{1}\,0.9^{9}=.3874$$

$$P(X=2)=\binom{10}{2}\,0.1^{2}\,0.9^{8}=.1937$$

P(X = 2) = .9298

Example:4.5.2.. What is the probability of getting no more than 2 misstated accounts in a simple random sample of size 10 drawn without replacement from an accounting population that has a 10% misstatement rate.

Population Size	**20**	**200**	**2000**	**Bin Approx**
P(X = 0)	.2368	.3398	.3476	.3487
P(X = 1)	.5263	.3974	.3881	.3874
P(X = 2)	.2368	.1975	.1939	.1937
P(X = 2)	**.9999**	**.9347**	**.9296**	**.9298**

Remark: Let $Y \sim B(m_1, \pi)$, $X \sim B(m_2, \pi)$. Then, the conditional distribution of Y given $X + Y = s_1$ is

$$P(Y = y \mid X + Y = s_1) = \frac{\binom{m_1}{y}\binom{m_2}{s_1 - y}}{\binom{m_1 + m_2}{s_1}}.$$

4.6. MEAN AND VARIANCE OF HYPERGEOMETIC DISTRIBUTION:

mean $E(x) = \sum_{k=0}^{n} K.\ P(x=k) = \sum_{k=0}^{n} K \left\{\binom{M}{K}\binom{N-M}{n-K} \div \binom{N}{n}\right\}$

$$= \frac{M}{\binom{N}{n}} \sum_{k=1}^{n} \left\{\binom{M-1}{K-1}\binom{N-M}{n-k}\right\}$$

$$= \frac{M}{\binom{N}{n}} \sum_{x=0}^{m} \binom{A}{x}\binom{N-A-1}{m-x}$$ where x=K-1, m = n-1
M-1 = A

$$= \frac{M}{\binom{N}{n}} \binom{N-1}{m}$$

$$= \frac{M}{\binom{N}{n}} \binom{N-1}{n-1}$$

$$= \frac{nM}{N}$$

Var $E[x(x-1)] = \sum_{k=0}^{n} K(K-1) \left\{\binom{M}{K}\binom{N-M}{n-K} \div \binom{N}{n}\right\}$

$$= \frac{M(M-1)}{\binom{N}{n}} \sum_{k=2}^{n} \left\{\binom{M-2}{K-2}\binom{N-M}{n-k}\right\}$$

$$= \frac{M(M-1)}{\binom{N}{n}} \binom{N-2}{n-2} = \frac{M(M-1)\ n(n-1}{N(N-1)}$$

$\therefore$ $E(x2) = E[x(x-1) + E(x) = \frac{M(M-1)\ n(n-1)}{N(n-1)} + \frac{nM}{N}$

Hence $V(x) = \frac{M(M-1)\ n(n-1)}{N(N-1)} + \frac{nM}{N} - \frac{(nM)^2}{N}$

$$\text{Var}(x) = \frac{n(m/N)(1-m/N)(N-n)}{(N-1)}$$

Example4.6.1. From a group of 20 Ph.D. engineers, ten are randomly selected for employment. Find the probability that the ten selected include all five ofbest engineers.

Solution. **N = 20**

n = 10

S = 5

(N-S) =15

X = 5

$$\textbf{P(X=5)} = \frac{\left({}^{5}C_{5}\right)\left({}^{5}C_{15}\right)}{{}^{10}C_{20}} \quad \frac{1 \text{ x } 3{,}003}{184{,}756} = \textbf{0.0162}$$

Remark: The Binomial as an Approximation to the Hypergeometric

Since the Hypergeometric is the Binomial distribution without replacement. How do they match up ?

Example4.6.2.:

Suppose N = 25, S = 15, n =2, (N-S) = 10 X = 0, 1 or 2

<u>Hypergeometric Solution:</u>

$$\textbf{P(X=0)} = \frac{\left({}^{0}C_{15}\right)\left({}^{2}C_{10}\right)}{{}^{2}C_{25}} = \frac{1\text{x}45}{300} = \textbf{0.15}$$

$$\textbf{P(X=1)} = \frac{\left({}^{1}C_{15}\right)\left({}^{4}C_{10}\right)}{{}^{2}C_{25}} = \frac{15\text{x}10}{300} = \textbf{0.50}$$

$$\textbf{P(X=2)} = \frac{\left({}^{2}C_{15}\right)\left({}^{0}C_{10}\right)}{{}^{2}C_{25}} = \frac{105\text{x}1}{300} = \textbf{0.35}$$

Binomial Solution:

p = Successes/Total - 15/25 = **0.6** So, (1-p) = **0.4, n = 2**

P(X=0) = ${}^{0}C_{2}(0.6)^{0}.(0.4)^{2} = (1)(1)(0.16)$ = **0.16**

P(X=1)= ${}^{1}C_{2}\ (0.6)^{1}.(0.4)^{1} = (2)(0.6)(0.4)$ **= 0.48**

P(X=2)= ${}^{2}C_{2}\ (0.6)^{2}.(0.4)^{0} = (1)(0.36)(1)$ = **0.36**

1.00

Remark: One should only use the Binomial to approximate the Hyper-geometric when "n" is large.

$$\cdot\left[\frac{N(N+1)-6N(N-n)}{m(N-m)}+\frac{3n(N-n)(n+6)}{N2}-6\right]$$

4.7 LESSON END EXERCISES

1) An urn contains ten (10) marbles of which five (5) are green, two (2) are blue, three (3) are red. Three marbles are to drawn from the urn, one at a time, without replacement. Show that the probability that all three marbles drawn are green is **(0.0833) or 1/12**

2) A warehouse contains ten printing machines, four of which are defective. A company selects five (5) of the machines at random, thinking that all are in working condition. Show that the probability that all five of the machines are non-defective is **(0.0238) or 1/42**

3) A jury of six (6) people was selected from a group of twenty (20) potential jurors, of whom eight (8) were African-America and twelve (12) were White. Show that the probability that the jury contains one or fewer African-Americans is **0.187**

4) A hat contains 20 names, 12 of which are females. If five names are drawn from the hat, what is the prob. that there are at least two female names drawn? Calculate the expected number?

5) A marble bag contains ten red and 15 green marbles. If Kelly reaches into bag and

withdraws five marbles, what is the probability that she will get

a. Exactly one red marble.

b. At least 2 red marbles.

c. No green marbles.

6) What is the probability that a bridge hand of 13 cards contains six spades, four hearts, two diamonds and one club?

7) A package contains ten yellow, six green, eight purple, and nine red candies jumbled together. What is the probability that at least three red candies are drawn among five candies chosen from the package? What is the expected number of red candies among five candies poured from the package?

4.8. SUGGESTED READINGS

1. Goon,Gupta;Das Gupta (1991).Fundamental of Statistics

2. Gupta,S.C. and Kapoor,V.K. Fundamental of Mathematical Statistics

3. Hoel,P.G. (1971).Introductory of Mathematical Statistics

4. Hogg R.V and Craig,A.T.Introduction to Mathematical Statistics

5. Hogg ,RV and Tanis,EA(1993).Probability and Statistical Inference

6. Mood,AM,Bose DC and Graybill F A.Introduction to the Theory of Statistics

7. Rohtagi,VK.An Introduction to Probability Theory and Mathematical Statistics

******

UNIT-II LESSON - 5

UNIFORM OR RECTANGULAR DISTRIBUTION

STRUCTURE

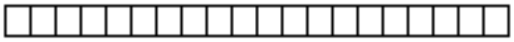
5.1 Objectives:

The main objectives of this chapter are

1. To introduce students to uniform distribution
2. To discuss utility and important properties of uniform distribution

5.2 Introduction:

A uniform distribution is one for which the probability of occurrence is the same for all values of X. It is sometimes called a rectangular distribution. For example, if a fair die is thrown, the probability of obtaining any one of the six possible outcomes is 1/6. Since all outcomes are equally probable, the distribution is uniform. If a uniform distribution is divided into equally spaced intervals, there will be an equal number of members of the population in each interval.



A uniform distribution

A nonuniform distribution

5.3 Definition:

A continuous random variable X is said to have Uniform or Rectangular Distribution if its p.d.f is given by

$$\int(x)=\begin{cases}\frac{1}{b-a} & a\le x\le b,(b>a)\\ \text{O, Otherwise}\end{cases}$$

The function defined above is called the p.d.f. of Uniform or Rectangular Distribution on the interval (a, b) and we write X~U (a, b) distribution. a and b are the parameters of the distribution. Here a is location parameter and b – a is scale parameter.

For uniform distribution, subintervals of (a, b) that have equal lengths, all have the same induced probability and hence all are equally likely. If a number is at random chosen in the interval (a, b) this means that a number is the observed value of X where X is U(a, b) that's why the distribution is called uniform distribution. The probability that X takes on value in the interval $(c,d)\subseteq(a,b)$ is given by

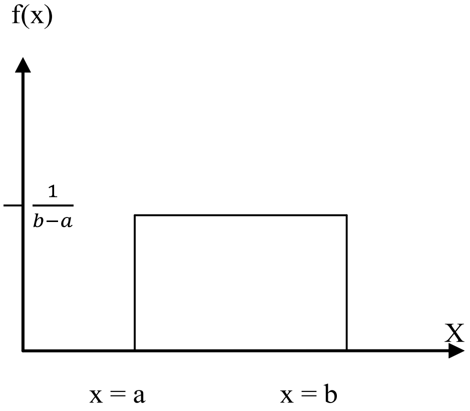
$$P = \int_c^d f(x)dx = \frac{d-c}{b-a}$$

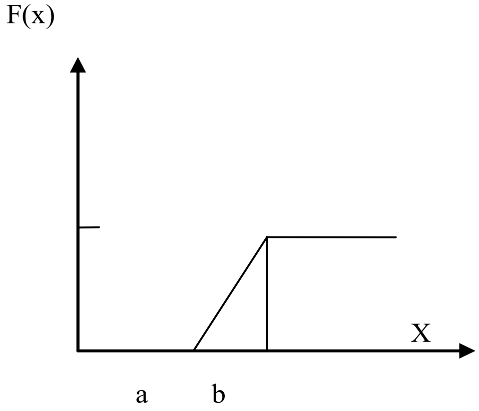
The distribution is also called rectangular distribution, since the graph of the p.d.f. represents a rectangular over the x-axis and between the ordinates $x = a$, and $x = b$

The distribution function of U(a, b)distributionis given by

$$F_x(x) = \begin{cases} 0 & x < a \\ \frac{N-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$

The graphic representation of $f(x)$ and $F_x(x)$ are given below





Some other forms of uniform distribution are

$$f(x) = \begin{cases} \frac{1}{2a}, -a \le x \le a \\ \text{O, otherwise} \end{cases}$$

This is denoted by U(- a,a)

$$f(\mathrm{x}) = \begin{cases} 1, & 0 \le \mathrm{x} \le 1 \\ 0, & \text{otherwise} \end{cases}$$

This is denoted by U(- a,a)

$$f(\mathrm{x}) = \begin{cases} 1, & 0 \le \mathrm{x} \le 1 \\ 0, & \text{otherwise} \end{cases}$$

This is denoted by U (0,1)

$$f(\mathrm{x}) = \begin{cases} \frac{1}{\theta}, -a \le x \le \theta \\ \mathrm{O}, \text{otherwise} \end{cases}$$

This is denoted by U(0, θ)

5.4 Moments of Uniform Distribution:

Let μ_r denotes the rth moment about origin then

$$\mu_r = E(X^r) = \int_a^b x^r f(x)dx$$

$$= \int_a^b \frac{\mathrm{x}^r}{\mathrm{b}-\mathrm{a}} \mathrm{dx} \qquad = \frac{1}{\mathrm{b}-\mathrm{a}}\left[\frac{b^{r+1}-a^{r+1}}{r+1}\right]$$

In particular,mean $\mu_1 = \frac{1}{\mathrm{b}-\mathrm{a}}\left[\frac{\mathrm{b}^2-\mathrm{a}^2}{2}\right] = \frac{\mathrm{b}+\mathrm{a}}{2}$

And $\mu_2 = \mathrm{E}(\mathrm{X}^2) = \frac{1}{\mathrm{b}-\mathrm{a}}\left[\frac{\mathrm{b}^3-\mathrm{a}^3}{2}\right] = \frac{1}{3}(\mathrm{b}^2+\mathrm{ab}+\mathrm{a}^2)$

Therefore variance of the distribution is

$$\mu_2 = \mathrm{E}(\mathrm{X}^2) - [\mathrm{E}(\mathrm{X})]^2 = \mu_2 - \mu_1^2$$

$$= \frac{1}{3}(\mathrm{b}^2+\mathrm{ab}+\mathrm{a}^2) - \left[\frac{\mathrm{b}+\mathrm{a}}{2}\right]^2 = \frac{1}{12}[\mathrm{b}-\mathrm{a}]^2$$

5.5 Moment Generating Function:

The Moment Generating Function (m.g.f) of the uniform distribution is given by

$$M_x(t) = E(e^{tx}) = \int_a^b e^{tx} f(x)dx$$

$$= \int_a^b e^{tx} \frac{1}{b-a} dx \qquad = \frac{e^{bt} - e^{at}}{t(b-a)}, t \neq 0$$

Expanding the exponential terms we get

$$M_x(t) = \frac{1}{b-a} \sum_{n=1}^{\infty} \left(\frac{b^n - a^n}{n!} \right) t^{n-1}$$

$$= \frac{1}{b-a} \sum_{n=0}^{\infty} \left(\frac{b^{n+1} - a^{n+1}}{n+1} \right) \frac{t^n}{n!}$$

Now we know that

$$\mu_n = \text{coefficient of } \frac{t^n}{n!} = \frac{1}{b-a} \left(\frac{b^{n+1} - a^{n+1}}{n+1} \right)$$

In particular, mean $\mu_1 = \frac{1}{b-a} \left[\frac{b^2 - a^2}{2} \right] = \frac{b+a}{2}$

$$\mu_2 = \frac{1}{b-a} \left[\frac{b^3 - a^3}{2} \right] = \frac{1}{3}(b^2 + ab + a^2)$$

$$\mu_2 - \mu_1^2$$

$$\mu_2 = \frac{1}{3}(b^2 + ab + a^2) - \left[\frac{b+a}{2} \right]^2 = \frac{1}{12}[b-a]^2$$

5.6 Mean Deviation about Mean:

Mean Deviation about Mean of uniform distribution is given by

$$\text{M.D.} = E(|X - \mu|) = \int_a^b |x - \mu|\, f(x)dx \qquad = \frac{1}{b-a}\int_a^b \left| x - \frac{b+a}{2} \right| dx$$

$$\frac{1}{b-a}\int_{-c}^{c} |z|\, dz, \qquad \left[Z = x - \frac{b+a}{2},\ c = \frac{b-a}{2} \right]$$

$$= \frac{2}{b-a}\int_0^c |z|\, dz, \qquad = \frac{c^2}{b-a} = \frac{b-a}{4}$$

6.7 Median:

Let m be the median of the distribution ,then we have

$$\frac{1}{2} = \int_a^m f(x)dx \qquad = \frac{1}{b-a}\int_a^m dx$$

$$= \frac{m-a}{b-a} \Rightarrow m = \frac{b+a}{2}$$

Thus for uniform distribution mean = median

5.8 Mode:

Since in the rectangle (b–a) $(b-a)^{-1}$, each point in the interval(a,b) has the maximum probability, so each point of (a,b) is mode .Some authors say that mode of uniform distribution does not exists.

5.9 Theorem:

If X is a random variable with continuous distribution function F, then F(X) has a uniform (0,1) distribution.

Proof: Since F is a distribution function, it is non decreasing. Let Y = F(X),then the distribution function G of Y is given by

$$G_y(y) = P(Y \le y) = P[F(X) \le y] = P[X \le F^{-1}(y)]$$

Inverse exists because F is give to be non decreasing and continuous

$$\therefore G_y(y) = F[F^{-1}(y)]$$

Since F is the distribution function of X.

$$\therefore G_y(y) = y$$

Therefore the p.d.f. of Y = F(X) is given by

$$g_y(y) \qquad = \frac{d}{dy} G_y(y) = 1$$

Since F is the distribution function , Y assumes values in the range (0,1)

Hence $g_y(y) \qquad = 1, 0 \le y \le 1$

Thus Y is uniform variate on (0,1)

5.10 Illustrations:

(1) If X is uniformly distributed over (0,10),calculate

(a) P(X<3) **(b)** P(X>7) **(c)** P(1<X<6)

Sol. Given that X~U (0, 10), so

$$f(x) = \frac{1}{10}, \quad 0 \le x \le 10$$

Now $P(X<3) = \int_0^3 f(x)dx \quad \int_0^3 \frac{1}{10}dx \qquad = \frac{3}{10}$

$$P(X>7) = \int_7^{10} \frac{1}{10}dx = \frac{3}{10}$$

$$P(1<X<6) = \int_7^6 \frac{1}{10}dx \qquad = \frac{5}{10} = \frac{1}{2}$$

(2) If X is uniformly distributed with mean 1 and variance 1/3 ,find P(X<0)

Sol. We know that if X~U(a, b), then

$$f(x)=\frac{1}{b-a}, a\le x\le b$$

And $\text{mean(X)}=\frac{a+b}{1}$, $\text{var(X)}=\frac{1}{12}(b-a)^2$

Thus, by given conditions $\frac{a+b}{1}=1, \frac{1}{12}(b-a)^2=\frac{1}{3}$

$\Rightarrow \text{a}+\text{b}=2,\ \text{b}-\text{a}=\pm 4$

Since b>a, solving these relations we get

$$\text{a}=-1, \text{b}=3$$

Hence $f(x)=\frac{1}{4}, \qquad -1\le x\le 3$

Therefore required probability is given by

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

(3) If X is uniformly distributed over [1,2], find z such that

$$\text{P}(\text{X}>\text{Z}+\mu)+=\frac{1}{4}, \text{where } \mu=\text{E(X)}$$

Sol. Since X~U(1, 2), so

$$f(x)=1, \qquad 1\le x\le 2$$

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

Given $\text{P}(\text{X}>\text{z}+\mu)=\frac{1}{4}\Rightarrow, \text{P}\left(\text{X}>\text{z}+\frac{3}{2}\right)=\frac{1}{4}$

$$\Rightarrow \int_{z+\frac{3}{2}}^{2} f(x)dx = \frac{1}{4} \qquad \Rightarrow \int_{z+\frac{3}{2}}^{2} dx = \frac{1}{4}$$

$$\Rightarrow 2-\left(z+\frac{3}{2}\right)=\frac{1}{4} \Rightarrow z=\frac{1}{4}$$

(4) Show that for the rectangular distribution

$$f(x)=\frac{1}{2a}, -a<x<a$$

The m.g.f. about origin is $\frac{1}{at}$ sing at. Also show that moments of even order are given by $\mu_{2n}=\frac{a^{2n}}{(2n+1)}$

Sol. The m.g.f.of a random variable about origin is given by

$$M_x(t)=E(e^{tx})=\int_{-a}^{a} e^{tx} f(x)dx$$

$$=\int_{-a}^{a} \frac{1}{2a} e^{tx} dx \quad = \frac{1}{2a}\left|\frac{e^{tx}}{t}\right|_{-a}^{a}$$

$$=\frac{1}{2a}(e^{at}-e^{-at})=\frac{\sinh at}{at}$$

$$=\frac{1}{at}\left[at+\frac{(at)^3}{3!}+\frac{(at)^5}{5!}+......\right]$$

$$=1+\frac{a^2t^2}{3!}+\frac{a^4t^4}{5!}+...........$$

Since there are no terms with odd powers of t in $M_x(t)$, all moments of odd orders about origin vanish, i.e.

$$\mu_{2n+1} \text{ (about origin)} = 0$$

In particular μ_1 (about origin) = 0 i.e. mean = 0

Thus μ_r (about origin) = μ_r

Hence $\mu_{2n+1} = 0$; n= 0,1,2 ……..

Thus all odd order moments about mean vanish. The moments of enen order are given by

$$\mu_{2n} = \text{coefficient of } \frac{t^{2n}}{n!} \text{ in } M_x(t) = \frac{a^{2n}}{(2n+1)}$$

(5) If X has uniform distribution on [0,1] ,find the distribution of Y=–2logX . Identify the distribution also.

Sol. Given Y = – 2logX Then the distribution function G of Y is given by

$$G_y(y) = P(Y \le y) = P(-2\log X \le y)$$

$$= P\left(\log \ge -\frac{y}{2}\right) = P\left(X \ge e^{-\frac{y}{2}}\right) = 1 - P\left(X \le e^{-\frac{y}{2}}\right)$$

$$= 1 - \int_0^{e-\frac{y}{2}} f(x)dx = 1 - \int_0^{e-\frac{y}{2}} 1.dx = 1 - e - \tfrac{y}{2}$$

Hence the p.d.f. of Y is given by

$$g_y(y) = \frac{d}{dy} G_y(y) = \frac{1}{2} e - \tfrac{y}{2}, 0 < Y < \infty$$

[Since as X ranges in (0,1) ,Y =–2logX ranges from 0 to ∞]

(6) Subway trains on a certain line run every half hour between midnight and six in morning .What is the probability that a man entering the station at random time during this period will have to wait at least twenty minutes?

Sol. Let the random variable X denotes the waiting time (in minutes)for the next train.Under the assumption that a man arrives at the station at random, X is uniformly distributed on (0,30) with p.d.f.

$$f(x)=\begin{cases}\dfrac{1}{30}, & 0<x<30\\ 0, & otherwise\end{cases}$$

The probability that he has to wait at least 20 minutes is given by

$$P(X\geq 20)=\int_{20}^{30}f(x)dx=\frac{1}{30}\int_{20}^{30}1.fdx$$

$$=\frac{1}{30}(30-20)=\frac{1}{3}$$

5.11 Summary:

In this lesson we have discussed an important continuous distribution namely Uniform distribution. Various forms of the distribution are presented and useful properties of uniform distribution have been proved, its relation with some other distributions has also been discussed. Illustrations are given through examples.

5.12 Self Assignment:

(1) A random variable X has a uniform distribution over (-3, 3),find

(i) P(X = 2) ,P(X< 2), P(|X| < 2) and P(|X-2| < 2)

(ii) find k for which P(X >k) = $\frac{1}{3}$

(2) Suppose X is uniformly distributed over $(-\alpha, \alpha), where\ \alpha > 0$.Determine α such that

(i) $P(X>1)=\dfrac{1}{3}$

(ii) $P(X>\dfrac{1}{2})=0.3$ and

(iii) $P(|X|<1)=P(|X|>1)$

(**3**) Two variates x and y are independently and uniformly distributed in the interval (0,6) and (0,9) respectively. Find the probability that the equation $x^2 - ax + b = 0$ has two real roots.

5.13 Further Reading:

1. Continuous Univariate Distributions,N.K.Johnson and s. Kotz
2. Introduction to the Theory of Statistics: A M Mood ,F A Graybill and D C Boes
3. Fundamentals of Mathematical Statistics : S C Gupta and V K Kapoor
4. New Mathematical Statistics : Bansi Lal and Sanjay Arora
5. Introduction to Probability Models : Sheldon M Ross

****

UNIT-II **LESSON-6**

NORMAL DISTRIBUTION AND ITS APPLICATIONS

STRUCTURE

6.1 OBJECTIVES

1. To introduce the students the concept of normal distribution.

2. To acquaint students with the importance of normal distribution.

3. To tell students how to use normal distribution with the help of examples.

6.2 INTRODUCTION:

Normal distribution was first discovered in 1733 by English mathematician De-Moivre, who obtained this continuous distribution as a limiting case of the binomial distribution and applied it to problems arising in the game of chance. It

Laplace, no later than 1774 but through a historical error it was credited to Gauss, who first made reference to it in the beginning of 19th century (1809), as the distribution of errors in Astronomy. The normal model has, nevertheless, become the most important probability model in statistical analysis.

6.3 Definition and P.d.f :

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

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

Remarks. 1. A random variable X with mean μ and variance σ^2 and following the normal law () is expressed by $X \sim N(\mu, \sigma^2)$. Here μ and σ^2 are two parameters, μ is known as location parameter and σ^2 as scale parameters.

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

3. The p.d.f. of standard normal variate Z is given by

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{z}}, -\infty < z < \infty$$

and the corresponding distribution function, denoted by $\Phi(z)$ is given by

$$\Phi(z) = P(Z \le -z) = \int_{-\infty}^{\infty} \varphi(u)du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du$$

6.4 Important results on the distribution function $\Phi(.)$ of SNV:

Result 1. $\Phi(-z) = 1 - \Phi(z)$

Proof. $\Phi(-z) = 1 - \Phi(z)$

$$\Phi(-z) = P(Z \le -z) = P(Z \ge z) \qquad \text{(By symmetry)}$$

$$= 1 - P(Z \le z)$$

$$= 1 - \Phi(z)$$

Result 2. $P(a \le X \le b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$, where $X \sim N(\mu, \sigma^2)$

Proof. $P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right); \qquad \left(Z = \frac{X-\mu}{\sigma}\right)$

$$= P\left(Z \le \frac{b-\mu}{\sigma}\right) - P\left(Z \le \frac{a-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

6.5 Normal Distribution as a Limiting form of Binomial Distribution:

Normal distribution is another limiting form of thr binomial distribution under the following conditions:

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

(ii) neither p nor q is very small.

The probability function of the binomial distribution with parameters n and p is given by

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}; = 0, 1, 2, \ldots., n \qquad \ldots(18.5.1)$$

Let us now consider the standard binomial variate :

$$Z = \frac{X - E(X)}{V(X)} = \frac{X - np}{\sqrt{npq}}; X = 0, 1, 2, \ldots\ldots, n \qquad \ldots(18.5.2)$$

When $\qquad X = 0, Z = \frac{-np}{\sqrt{npq}} = \sqrt{{np}/{q}}$

And when $\quad X = n, Z = \frac{n - np}{\sqrt{npq}} = \sqrt{{np}/{q}}$

Thus in the limit as $n \to \infty$, Z takes the values from $-\infty$ to ∞. Hence the distribution of X will be a continuous distribution over the range $-\infty$ to ∞.

We want the limiting form of (18.5.1) under the above two conditions. Using Stirling's approximation to r! for large r, viz.,

$$\lim_{r\to\infty} r! \cong \sqrt{2\pi} e^{-r} r^{r+\left(\frac{1}{2}\right)},$$

We have in the limit as $n \to \infty$ and consequently $\to \infty$,

$$\lim p(x) = \lim\left[\frac{\sqrt{2\pi} s^{-n} n^{n+\left(\frac{1}{2}\right)} p^{x} q^{n-x}}{\sqrt{2\pi} s^{-x} x^{x+\left(\frac{1}{2}\right)} \sqrt{2\pi} s^{-(n-x)} (n-x)^{n-x+\frac{1}{2}}}\right]$$

$$= \lim\left[\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}}, \frac{(np)^{x+\frac{1}{2}} (nq)^{n-x+\frac{1}{2}}}{(n)^{x+\frac{1}{2}} (n-x)^{n-x+\frac{1}{2}}}\right]$$

$$= \lim\left[\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}} . \left(\frac{np}{x}\right)^{x+\frac{1}{2}} \left(\frac{nq}{n-x}\right)^{n-x+\frac{1}{2}}\right] \qquad \ldots(18.5.3)$$

From (18.5.2), we have

$$X = np + Z\sqrt{npq} \qquad \Rightarrow \frac{x}{np} = 1 + Z\sqrt{{q}/{(np)}}$$

Also

$$n - X = n - np - Z\sqrt{npq} = np - Z\sqrt{npq}$$

$$\therefore \frac{n\text{ - }x}{nq} = 1 - Z\sqrt{{q}/{(np)}}. \text{ Also } dz = \frac{1}{\sqrt{npq}} dx$$

Hence the probability differential of the distribution of Z, in the limit is given from (18.5.3) by d G(z)= g(z)= $\lim_{n\to\infty}\left[\frac{1}{2\pi}\,x\,\frac{1}{N}\right]$dz ……… ….(18.5.4)

where $N=\left[\frac{x}{np}\right]^{x+\frac{1}{2}}\left[\frac{n-x}{nq}\right]^{n-x+\frac{1}{2}}$

$$\log N=\left(x+\frac{1}{2}\right)\log\left(x\big/_{np}\right)+\left(n-x+\frac{1}{2}\right)\log\left\{(n-x)\big/_{nq}\right\}$$

$$=\left(np+z\sqrt{npq}+\frac{1}{2}\right)\log\left[1+z\sqrt{\left(q\big/_{np}\right)}\right]$$

$$+\left(np-z\sqrt{npq}+\frac{1}{2}\right)\log\left[1-z\sqrt{\left(p\big/_{np}\right)}\right]$$

$$=\left(np+z\sqrt{npq}+\frac{1}{2}\right)\log\left[1+z\sqrt{\left(q\big/_{np}\right)}\right]\left[z\sqrt{\left(q\big/_{np}\right)}-\frac{1}{2}z^2\left(q\big/_{np}\right)+\frac{1}{3}z^3\left(q\big/_{np}\right)^{3\!/\!2}-\ldots\right]$$

$$+\left(nq-z\sqrt{npq}+\frac{1}{2}\right)\left[-z\sqrt{\left(p\big/_{np}\right)}-\frac{1}{2}z^2\left(p\big/_{nq}\right)-\frac{1}{3}z^3\left(p\big/_{nq}\right)^{3\!/\!2}-\ldots\right]$$

$$=z\sqrt{npq}-\frac{1}{2}qz^2+\frac{1}{3}z^3\frac{q^{\frac{3}{2}}}{\sqrt{np}}+z^3q-\frac{1}{2}z^3\frac{q^{\frac{3}{2}}}{\sqrt{np}}+\frac{1}{2}z\sqrt{q\big/_{np}}\quad-\frac{1}{4}z^2\frac{q}{np}+\ldots)+$$

$$(-z\sqrt{npq}-\frac{1}{2}z^2p-\frac{1}{3}z^3\frac{p^{\frac{3}{2}}}{\sqrt{nq}}+z^2p+\frac{1}{2}z^3\frac{p^{\frac{3}{2}}}{\sqrt{nq}}-\frac{1}{2}z\sqrt{p\big/_{nq}}-\frac{1}{4}z^2\frac{p}{nq}+\ldots\ \}]$$

i.e.,

$$\log N=\left[-\frac{1}{2}z^2(p+q)+z^2(p+q)+\frac{z}{2\sqrt{n}}\left\{\sqrt{\frac{p}{q}}+\sqrt{\frac{p}{q}}\right\}+0\left\{n^{-\frac{1}{2}}\right\}\right]$$

$$=\frac{z}{2}+o\left(n^{-\frac{1}{2}}\right)\rightarrow\frac{z^2}{2}\text{ as } n\rightarrow\infty$$

$$\therefore \lim_{n\to\infty}\log N = \frac{z^2}{2} \quad \Rightarrow \lim_{n\to\infty} N = e^{\frac{z^2}{2}}$$

Substituting in (18.5.4), we get

$$dG(z)= g(z)dz = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz, -\infty < z < \infty \qquad \text{............(18.5.4 a)}$$

Hence the probability function of Z is

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}, -\infty < z < \infty \qquad \text{............(18.5.4 b)}$$

This is the probability density function of the normal distribution with mean 0 and unit variance.

If X is a normal variate with mean μ and s.d. σ then $Z = \frac{x - \mu}{\sigma}$ is a standard normal variate. Jacobian of transformation is $1/\sigma$. Hence substituting in

(18.5.4 b), the p.d.f. of a normal variate X with E(X)= μ and Var(X)= σ^2 is given by

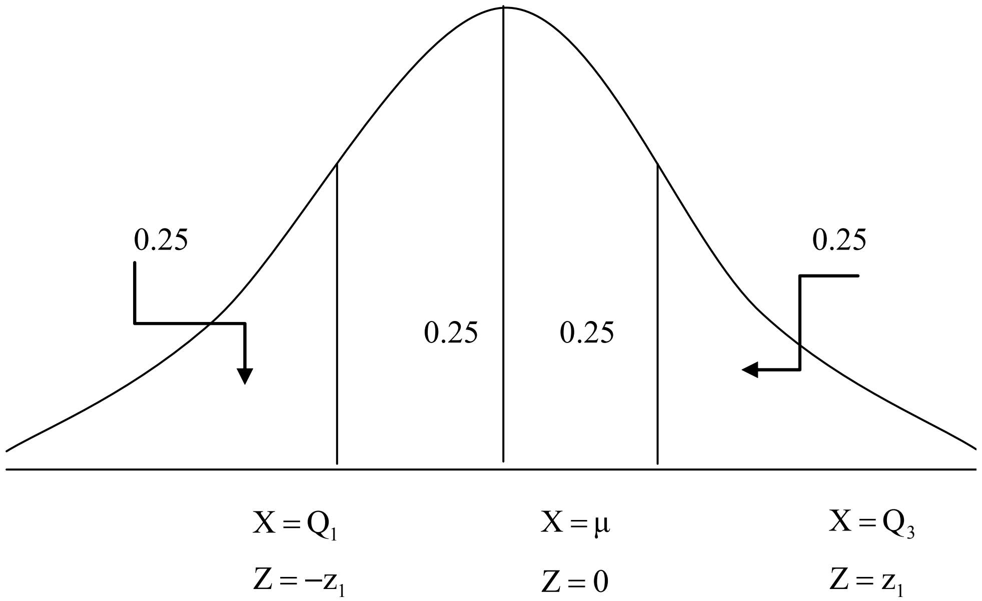
$$f_x(X) = \begin{matrix} \frac{1}{\sigma\sqrt{2\pi}} \exp[-{(x-\mu)^2}/{2\sigma^2}; -\infty < \mu < \infty, \sigma^2 < 0 \\ 0, \qquad \text{othewise} \end{matrix}$$

6.6 Theorem: Show that for a normal distribution quartile deviation ,mean deviation and standard deviation are approximately 10:12:15

Proof: Let X~N (μ, σ^2). If Q_1 and Q_3 are the first and the third quartiles,the by definition

P(X< Q_1)=0.25 and P(X > Q_3)=0.25

The points Q_1 and are Q_3 located as shown in the fig.below.



When X=$Q_3, Z = \dfrac{Q_3 - \mu}{\sigma} = z_1$(say)

And when $X = Q_1, Z = \dfrac{Q_1 - \mu}{\sigma} = z_1$ (by symmetry of normal curve)

Subtracting, we have

$\dfrac{Q_3 - Q_1}{\sigma} = 2z_1$, so that quartile deviation is given by

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \sigma z_1$$

From the fig. we have

$$P(0 < Z < z_1) = 0.25 \Rightarrow z_1 = 0.67 \text{ [from normal table]}$$

$$\therefore \text{Q.D.} = \sigma z_1 = 0.67\sigma \approx \frac{2}{3}\sigma$$

For normal distribution mean deviation about mean is given by

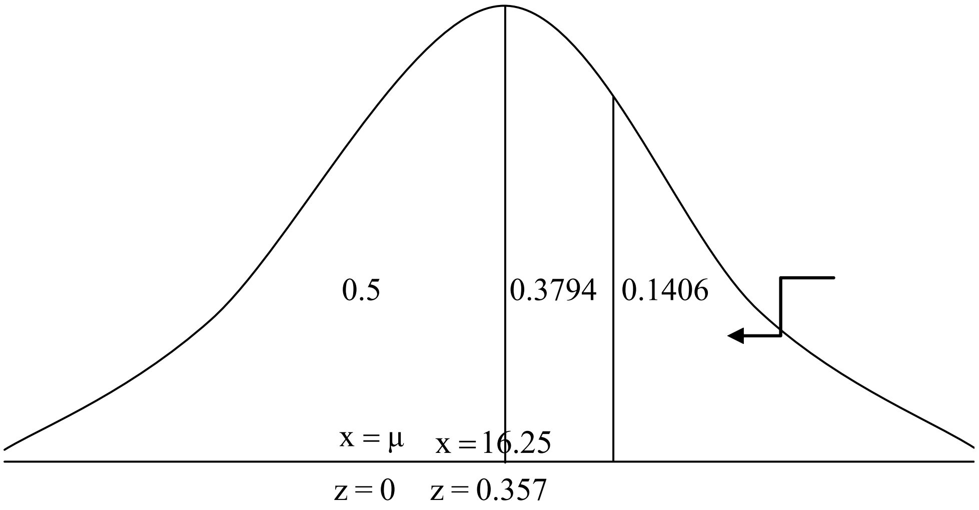
$$\text{M.D.} = \sigma\sqrt{2/\pi} \approx \frac{4}{5}\sigma$$

Hence Q.D.:M.D.:S.D. $:: \dfrac{2}{3}\sigma : \dfrac{4}{5}\sigma : \sigma :: \dfrac{2}{3} : \dfrac{4}{5} : 1 :: 10 : 12 : 15$

6.7 Illustrations:

(1) In a normal population with mean 15 and standard deviation 3.5, It is known that 647 observations exceed 16.25. What is the total no. of observations in the population.

Sol: Let n be the total no. of observation in the population and X be the random variable which represents the characteristics studied of the population then $X \sim N(15, (3.5)^2)$



It is given that

$$n.p\ (X > 16.25) = 647$$

$$\Rightarrow n.P\left(\frac{x-15}{3.5} > \frac{16.25-15}{3.5}\right) = 647$$

$\Rightarrow n.P\ (Z > .357) = 647$, where Z is a standard normal variate.

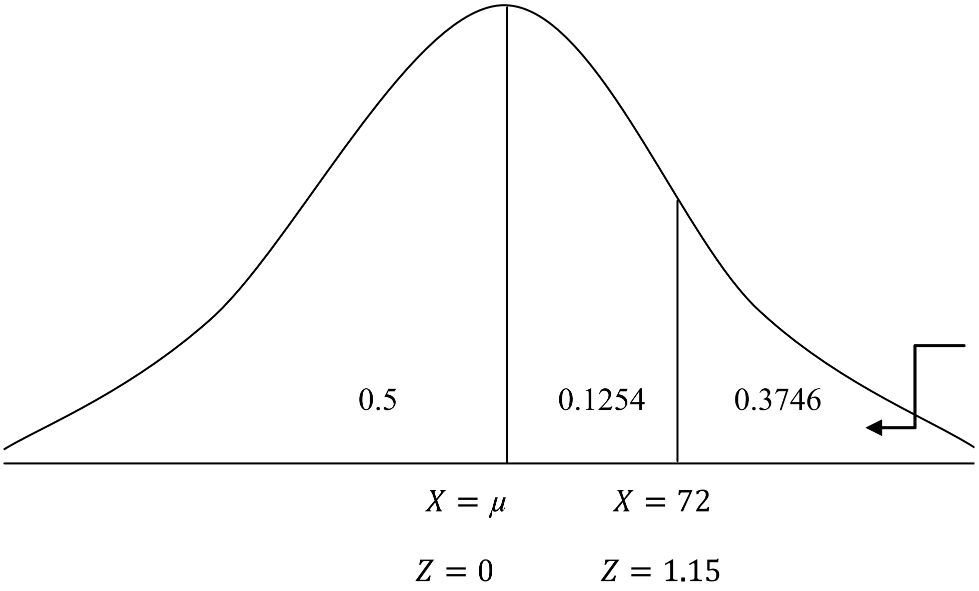
$$\Rightarrow n\,[0.5 - P(0 < Z < .357)] = 647$$

$$\Rightarrow n\ (0.5 - .1406 = 647) \Rightarrow n = \frac{647}{0.3594} \cong 1800$$

(2) Assume the mean height of the soldiers to be 68.22 inches with a variance of $10.8(\text{inch})^2$. How many soldiers in a regiment of 1000 would you expect to be over 6 feet tall?(Given that the area under the standard normal curve between Z=0 and Z=0.35 is 0.1368 and between Z=0 and Z=1.15 is 0.3746)

Sol: Let X be a random variable denoting the height of soldiers.

then X~N (68.22, 10.8)



let p be the probability that a soldier is over 6 feet tall i.e.,

$$p = P(X \geq 72) = P\left(\frac{x - 68.22}{\sqrt{10.8}} \geq \frac{72 - 68.22}{3.29}\right)$$

$$\Rightarrow p = P(Z \geq 1.15) = 0.5 - (0 \leq Z \leq 1.15)$$

$$= 0.5 - 0.3746 = 0.2154$$

Now the number of soldiers in a regiment of 1000 who are over 6 feet tall =1000 x p $\cong$ 125.

(3) For a certain normal distribution the first moment about 10 is 40 and fourth moment about 50 is 48. Find the mean and standard deviation of the distribution.

Sol: We know that if μ_1 is the first moment about the point X=A then arithmetic mean is given by :

Mean $= A + \mu_1$.

We are given that μ_1 (about the point X=10) = 40

$\Rightarrow$ Mean $= 10 + 40 = 50$

Also we are given μ_4 (about the point X = 50) = 48 i.e., $\mu_4 = 48$

But for a normal distribution with standard deviation 'σ'

$$\mu_4 = 3\sigma^4 \Rightarrow 3\sigma^4 = 48 \Rightarrow \sigma = 2$$

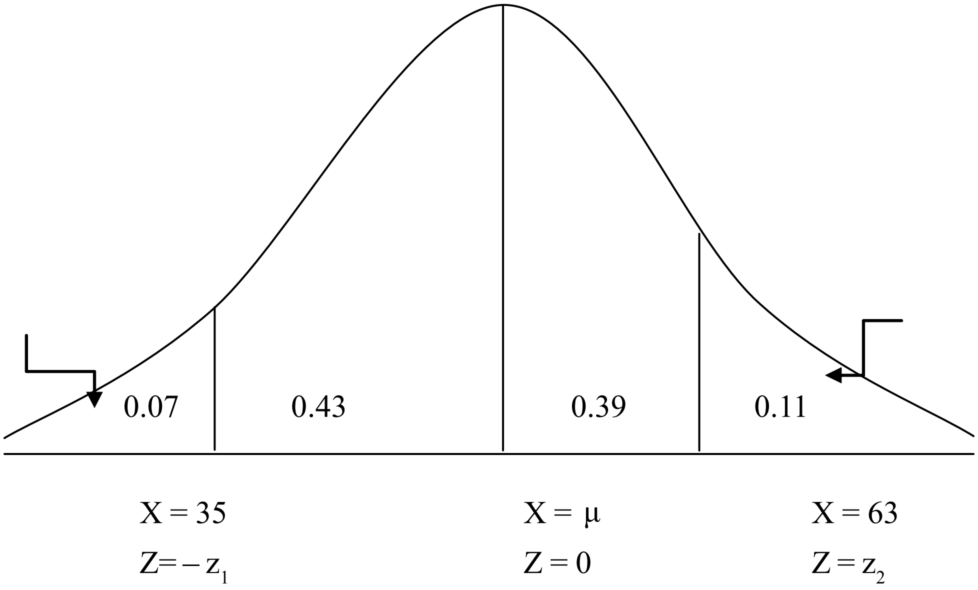
(4) In a distribution exactly normal, 7% of the items under 35 and 89% are under 63. Find the mean and standard deviation of the distribution.

Sol: Let X be a normal variate with mean μ and variance σ^2 then we are given that $P(X<63) = 0.89$ 0.07

$$\Rightarrow P(X > 63) = 0.11 \text{ and } P(X < 35) = 0.07$$

when $X=35, Z=\frac{35-\mu}{\sigma} = -z_1$, (say),

and when $X=63, Z=\frac{63-\mu}{\sigma} = -z_2$, (say),



Thus we have ,

$P(0<Z<z_2) = 0.39$ and $P(0<Z<z_1) = 0.43$

Hence from the normal tables, we have

$$z_2 = 1.23 \text{ and } z_1 = 1.48$$

$$\therefore \frac{63-\mu}{\sigma} = 1.23 \text{ and } \frac{35-\mu}{\sigma} = -1.48$$

Subtracting we get

$$\frac{28}{\sigma} = 2.71 \Rightarrow \sigma = \frac{28}{2.71} = 10.33$$

$$\therefore \qquad \mu = 35 + 1.48 \text{ x } 10.33 = 35 + 15.3 = 50.3$$

(5) If the skulls are classified as A, B and C according as the length-breadth index is under 75 between 75 and 80 or over 80. Find approximately (assuming that the distribution is normal). The mean and standard deviation of the series in which A are 58%. B are 38% and C are 4%. Given that if

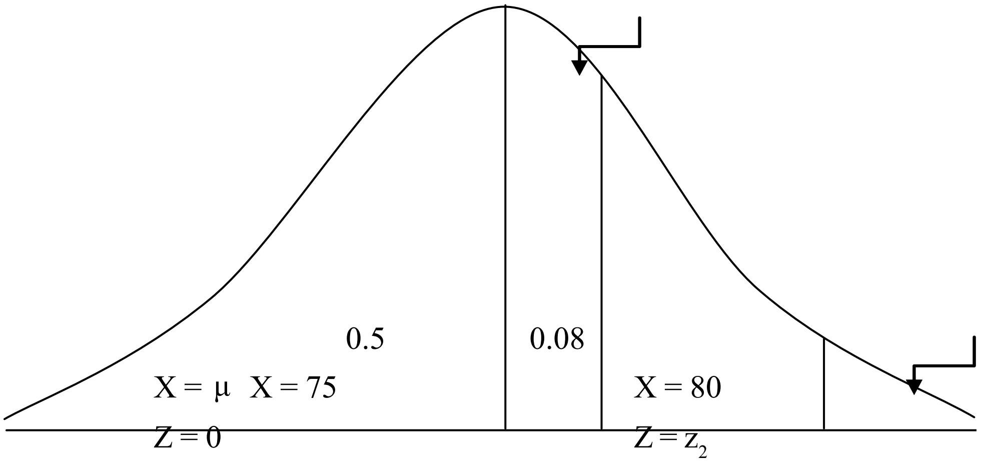
$$f(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{\left(\frac{-x^2}{2}\right)} dx,$$

then F(0.20) = 0.08 and f(1.75) = 0.46

Sol: Let the length-breadth index be denoted by the variable X , then we are given:

P(X < 75) = 0.58 and P(X >80) = 0.04 ....(1)

Since P(X < 75) represents the total area to the left of the ordinate at point X=75 and P(X>80) represents the total area to the right of the ordinate at point X=80. It is obvious that the points X=75 and X=80 are located at the positions shown in the figure below:



Now $f(t)=\frac{1}{\sqrt{2\pi}}\int_0^t e^{\left(\frac{-x^2}{2}\right)}dx$ represents the area under standard normal curve between the ordinates Z=0 and Z=t, Z being a N(0, 1) variate.

Hence $f(t)=\frac{1}{\sqrt{2\pi}}\int_0^t e^{\left(\frac{-x^2}{2}\right)}dx = P(0<Z<t)$

$\therefore$ $f(0.20)=P(0<Z<0.20)=0.08$ ……..(2)

and $f(1.75)=P(0<Z<1.75)=0.46$

Let μ and σ be the mean and standard deviation of the distribution then X~N (μ, σ^2)

When X=75, $Z=\frac{75-\mu}{\sigma}=z_1$ (say),

When X=80, $Z=\frac{80-\mu}{\sigma}=z_2$ (say),

From the figure it is obvious that

$P(X<75)=0.58 \Rightarrow P(0<Z<z_1)=0.08$

Using (2) we have $z_1=\frac{75-\mu}{\sigma}=0.20$ …………….(3)

Also $P(X<80)=0.04 \Rightarrow P(0<Z<z_2)=0.46$

From (2) we get $z_2=\frac{80-\mu}{\sigma}=1.75$ ……………….(4)

Solving (3) and (4) we get $\mu=74.4$ and $\sigma=3.2$

6.8 Summary:

In this section we have discussed the most important distribution viz. normal distribution which finds applications in almost every field of life as well as in every subject of study like Sociology, Economics, Medicine, Biology, Psychology etc.Normal distribution is derived as a limiting case of Binomial distribution. The application of normal distribution in various problems has been explained through examples.

6.9 Self Assessments:

1) Of a large group of men, 5% are under 60 inches in height and 40% are between 60 and 65 inches. Assuming the normal distribution find the mean height and standard deviation.

2) The marks obtained by a number of students for a certain subject are assumed to be normally distributed with mean 65 and standard deviation 5. If three students are taken at random from this set what is the probability that exactly two of them will have marks over 70.

3) In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and variance of the distribution. Given that the area between mean ordinates at any σ distance from mean,

$Z = \frac{x - \mu}{\sigma}$	: 0.496	1.405
Area	: 0.19	0.42

6.10 Further Readings:

1. Continuous Univariate Distributions,N.K.Johnson and s. Kotz
2. Introduction to the Theory of Statistics: A M Mood ,F A Graybill and D C Boes
3. Fundamentals of Mathematical Statistics : S C Gupta and V K Kapoor
4. New Mathematical Statistics : Bansi Lal and Sanjay Arora
5. Introduction to Probability Models : Sheldon M Ross

****

UNIT-II LESSON-7

NORMAL DISTRIBUTION AND ITS PROPERTIES

STRUCTURE

7.1 OBJECTIVES:

The main objectives are

1. To introduce normal distribution and its important features
2. To study its important properties
3. To discuss importance of normal distribution

7.2 INTRODUCTION

Normal distribution is the most important continuous distribution in the theory of Statistics due to a wide range of applications. Almost all the phenomenon occurring in the real world scenario can be studied with the help of normal distribution. Almost all the distributions occurring in practice can be approximated by normal distribution.

7.3 NORMAL DISTRIBUTION & ITS CHARACTERISTICS:

Definition: A r.v. X is said to have a normal distribution with parameters μ ,(called 'mean') and σ^2 (called 'variance') if its p.d.f. is given by the probability law

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

Here μ and σ^2 are called its parameters and we write it as $X\sim N\ (\mu,\sigma^2)$

Standard Normal Variate. If $X\sim N\ (\mu,\sigma^2)$, then $Z=\dfrac{X-\mu}{\sigma}$ is a standard normal variate with E(Z) =0 and Var (Z) = 1 and we write Z ~ N (0,1),

Proof: If $X\sim N\ (\mu,\sigma^2)$, and $Z=\dfrac{X-\mu}{\sigma}$ then

$$E(Z)=E\left[\frac{X-\mu}{\sigma}\right]=\frac{1}{\sigma}E[x-\sigma]=0$$ and

$$\text{Var(Z)=}\ \mathrm{Var}(Z)=V\left[\frac{x\text{ - }\mu}{\sigma}\right]=\frac{1}{\sigma^2}V[x-\sigma]=\frac{1}{\sigma^2}.\sigma^2=1$$

The p.d.f of standard normal variate Z is given by

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}; \qquad -\infty < z < \infty$$

Chief Characteristics of the Normal Distribution and Normal Probability Curve:
The p.d.f of normal distribution with mean μ and standard deviation σ is given by the equation:

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

and has the following properties

7.4 THE CURVE IS BELL-SHAPED AND SYMMETRICAL ABOUT THE LINE X = μ.

(ii) Mean, median and mode of the distribution coincide.

(iii) As x increases numerically, f(x) decreases rapidly, the maximum probability occurring at the point x = μ and is given by: $[p(x)]_{\max} = \dfrac{1}{\sigma\sqrt{2\pi}}$

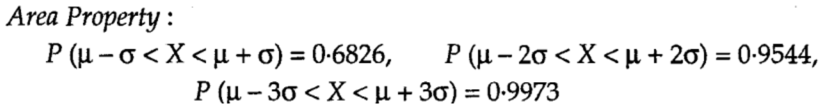
(iv) β_1=0and β_2=3.

(v) μ_{2r+1} =0,(r=0,1,2,...), i.e. odd order moments vanish

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

(vii) Linear combination of independent normal variates is also a normal variate.

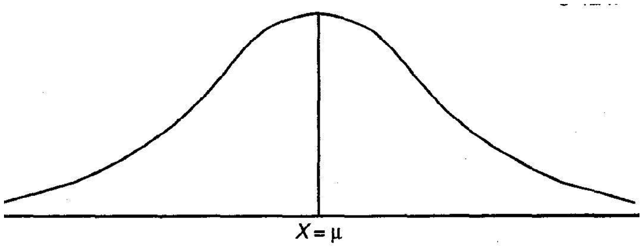
(viii)



Area Property :

$P(\mu - \sigma < X < \mu + \sigma) = 0{\cdot}6826, \qquad P(\mu - 2\sigma < X < \mu + 2\sigma) = 0{\cdot}9544,$

$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0{\cdot}9973$



7.5 MEAN OF NORMAL DISTRIBUTION:

If X~ N (μ , σ^2), we have

E[X] = Mean= $\int_{-\infty}^{\infty} xf(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}dx$ ……..(19.4.1)

Put Z = $\frac{x-\mu}{\sigma}$ so that $dx = \sigma\, dz$ and $x = \mu + \sigma z$ we get

$$\text{Mean}=\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu+\sigma z)\exp\left\{-\frac{1}{2}z^2\right\}\sigma\, dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu+\sigma z)\exp\left\{-\frac{1}{2}z^2\right\}dz$$

$$\frac{2\mu}{\sqrt{2\pi}} \int_{0}^{\infty} \exp\left\{-\frac{1}{2}z^2\right\}dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\exp\left\{-\frac{1}{2}z^2\right\}dz$$

$$=\frac{2\mu}{\sqrt{2\pi}} \int_{0}^{\infty} \exp\left\{-\frac{1}{2}z^2\right\}dz + 0 \qquad \text{…………(19.4.2)}$$

Since the integrand $z\exp\left\{-\frac{1}{2}z^2\right\}$ is an odd function of z

Now in $\int_{0}^{\infty} \exp\left\{-\frac{1}{2}z^2\right\}dz$, if we put

$$\frac{z^2}{2} = U \qquad \Rightarrow z = \sqrt{2U} \;\; and \;\; dz = \frac{1}{\sqrt{2U}}dU$$

We get $\int_{0}^{\infty} \exp\left\{-\frac{1}{2}z^2\right\}dz = \int_{0}^{\infty} e^{-U} \frac{dU}{\sqrt{2U}} = \frac{1}{\sqrt{2}} \int_{0}^{\infty} e^{-U} U^{\frac{1}{2}-1} dU$

$$= \frac{1}{\sqrt{2}}\Gamma\frac{1}{2} = \frac{1}{\sqrt{2}}\sqrt{\pi} = \sqrt{\frac{\pi}{2}}$$

Substituting in (19.4.2) we get

Mean= $\frac{2\mu}{\sqrt{2\pi}}\sqrt{\frac{\pi}{2}} = \mu$ Hence mean is μ

7.5 VARIANCE OF NORMAL DISTRIBUTION:

If X~ N (μ , σ^2), we have

$$\mu_2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-\mu)^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad \text{……..(19.5.1)}$$

Put Z = $\frac{x-\mu}{\sigma}$ so that $dx = \sigma dz$ and $x - \mu = \sigma z$ in (19.5.1) we get

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma z)^2 e^{-\frac{1}{2}(z)^2}\sigma dz = \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}(z)^2} dz \qquad \text{……….(19.5.2)}$$

put $\frac{z^2}{2} = U \qquad \Rightarrow z^2 = 2U \ and \ dz = \frac{1}{\sqrt{2U}}dU$ in (19.5.2) we get

$$\text{Variance}=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} 2Ue^{-U}\frac{dU}{\sqrt{2U}}$$

$$= \frac{\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-U}U^{\frac{1}{2}}dU = \frac{2\sigma^2}{\sqrt{\pi}}\int_{0}^{\infty} e^{-U}U^{\frac{1}{2}}dU$$

$$= \frac{2\sigma^2}{\sqrt{\pi}}\int_{0}^{\infty} e^{-U}U^{\frac{3}{2}-1}dU$$

$$= \frac{2\sigma^2}{\sqrt{\pi}}\Gamma\frac{3}{2} = \frac{2\sigma^2}{\sqrt{\pi}}\Gamma\frac{1}{2}+1 = \frac{2\sigma^2}{\sqrt{\pi}}\frac{1}{2}\sqrt{\pi} = \sigma^2$$

Hence Variance of Normal Distribution σ^2

7.6 MODE OF NORMAL DISTRIBUTION:

Mode is the value of x for which f(x) is maximum, i.e., mode is the solution of

$f'(x) = 0$ and $f''(x) < 0$

For normal distribution with mean with mean μ and standard deviation σ,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$ so that

$$\log f(x) = c - \frac{1}{2\sigma^2}(x-\mu)^2$$

where c = log ($\frac{1}{\sigma\sqrt{2\pi}}$), is a constant. Differentiating w.r. to x, we get

$$\frac{1}{f(x)}.f'(x) = -\frac{1}{2\sigma^2}(x-\mu)$$

$$\Rightarrow f'(x) = - \quad \frac{1}{2\sigma^2}(x-\mu)\, f(x)$$

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

$$-\frac{f(x)}{\sigma^2}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right] \ldots\ldots\ldots(19.6.1)$$

$$f'(x) = 0 \Rightarrow x - \mu = 0 \Rightarrow x = \mu$$

At $x = \mu$

$$f''(x) \quad = \quad -\frac{1}{\sigma^2}\left[f(x)\right]_{x=\mu} = -\frac{1}{\sigma^2}.\frac{1}{\sigma\sqrt{2\pi}} < 0$$

Hence x = μ is the mode of the normal distribution.

7.7 MEDIAN OF NORMAL DISTRIBUTION:

If M is the median of the normal distribution, we have

$$\int_{-\infty}^{M} f(x)dx = \frac{1}{2} \Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{M} e^{\left\{-(x-\mu)^2 \big/ 2\sigma^2\right\}} dx = \frac{1}{2}$$

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

But $\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} e^{\left\{-(x-\mu)^2 / 2\sigma^2\right\}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{\left(-z^2/2\right)} dz = \frac{1}{2}$

So we have

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

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} e^{\left\{-(x-\mu)^2 / 2\sigma^2\right\}} dx = 0, \text{ i.e., } \mu = M$$

Hence, for the normal distribution, Mean = Median.

7.8 M.G.F. OF NORMAL DISTRIBUTION:

If X~ N (μ , σ^2), then by definition m.g.f is given by

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

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

{ By substituting Z = $\frac{x-\mu}{\sigma}$ so that $dx = \sigma\, dz$ and $x = \mu + \sigma z$ }

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

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

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

$$=e^{\mu t+\frac{\sigma^2}{2}t^2}\times\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{u^2}{2}}\,du$$

Hence $M_X(t) = e^{\mu t+\frac{\sigma^2}{2}t^2}$

7.9 M.G.F. OF STANDARD NORMAL VARIATE:

If X~ N (μ , σ^2), then standard normal variate

given by $Z = \frac{x-\mu}{\sigma}$

$$M_z(t)=e^{-\mu t/\sigma}M_X\left(t/\sigma\right)=e^{-\mu t/\sigma}.e^{\left\{\frac{\mu t}{\sigma}+\left(t^2/\sigma^2\right)\sigma^2/2\right\}}$$

$$=e^{\frac{t^2}{2}}$$

Alternatively ,since if Z is a standard normal variate then $Z\sim N(0,1)$ .Hence taking $\mu=0$ *and* $\sigma^2=1$ in the m.g.f $M_X(t)$ of random variable X we get m.g.f of a standard normal variate.

7.10 LINEAR COMBINATION OF INDEPENDENT NORMAL VARIATES :

Let X_i (i = 1,2, 3, ..., n) be n independent normal variates with mean μ and variance σ^2 respectively. Then

$$\mathrm{M}_{X_i}(\mathrm{t}) = e^{\left\{\mu_i t + \left(t^2 \frac{\sigma_i^2}{2}\right)\right\}} \text{....................................}(19.10.1)$$

The m.g.f. of their linear combination $\sum_{i=1}^{n} a_i x_i$,where $a_1, a_2, ..., a_n$ are constant is given by

$$M_{\sum_{i=1}^{n} a_i X_i} = \prod_{i=1}^{n} M_{a_i X_i}(\mathrm{t}) \quad (\text{since } \mathrm{X_i^{'s}} \text{ are independent})$$

$$= M_{X1}(a_1 t) M_{X2}(a_2 t) \text{...........} M_{Xn}(a_n t) \qquad [\text{since } M_{cX}(t) = M_X(ct)]$$

From (19.10.1), we have

$$M_{Xi}(a_i t) = e^{\left\{\mu_i a_i t + \left(t^2 a_i^2 \frac{\sigma_i^2}{2}\right)\right\}}$$

$$M_{\sum_{i=1}^{n} a_i X_i}(t) = \left\{ e^{\mu_1 a_1 t + t^2 a_1^2 \sigma_1^2 / 2} \times e^{\mu_2 a_2 t + t^2 a_2^2 \sigma_2^2 / 2} \times \text{........} e^{\mu_n a_n t + t^2 a_n^2 \sigma_n^2 / 2} \right\}$$

$$= e^{\left[\left(\sum_{i=1}^{n} a_i \mu_i\right) t + t^2 \left(\sum_{i=1}^{n} a_i^2 \sigma_i^2\right) / 2\right]}$$

which is the m.g.f of a normal variate with mean $\sum_{i=1}^{n} a_i \mu_i$ and variance $\sum_{i=1}^{n} a_i^2 \sigma_i^2$.

Hence by uniqueness theorem of m.g.f,

$$\sum_{i=1}^{n} a_i X_i \sim N \left[\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right]$$

- If we take $a_1 = a_2 = 1, a_3 = a_4 = \ldots = 0$, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- If we take $a_1 = 1, a_2 = -1, a_3 = a_4 = \ldots = 0$, then $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$

Thus we see that the sum as well as the difference of two independent normal variates is also a normal variate.

- If we take $a_1 = a_2 = \ldots = a_n = 1$, then we get $\sum_{i=1}^{n} X_i \sim N\left[\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right]$

i.e., the sum of independent normal variates is also a normal variate, which establishes the additive property of the normal distribution.

7.11 MOMENTS OF NORMAL DISTRIBUTION:

Odd order moments about mean are given by

$$\mu_{2n+1} = \int_{-\infty}^{\infty} (x-\mu)^{2n+1} f(x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^{2n+1} \exp\{-(x-\mu)^2/2\sigma^2\} dx$$

$$\therefore \quad \mu_{2n+1} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^{2n+1} \exp(-z^2/2)\, dz,$$

$$= \frac{\sigma^{2n+1}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n+1} \exp(-z^2/2)\, dz = 0,$$

since the integrand $z^{2n+1} e^{-z^2/2}$ is an odd function of z.

Hence odd order moments of N.D vanish

Even order moments about mean are given by:

$$\mu_{2n} = \int_{-\infty}^{\infty} (x-\mu)^{2n} f(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^{2n} \exp(-z^2/2)\, dz$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n} \exp(-z^2/2)\, dz = \frac{\sigma^{2n}}{\sqrt{2\pi}} . 2 \int_{0}^{\infty} z^{2n} \exp(-z^2/2)\, dz$$

$$= \frac{2\sigma^{2n}}{\sqrt{2\pi}} \cdot \int_{0}^{\infty} (2t)^n e^{-t} \frac{dt}{\sqrt{2t}}, \qquad \left(t = \frac{z^2}{2}\right)$$

$$\therefore \quad \mu_{2n} = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \int_{0}^{\infty} e^{-t}\, t^{(n+\frac{1}{2})-1}\, dt \quad \Rightarrow \quad \mu_{2n} = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} . \Gamma\left(n + \frac{1}{2}\right)$$

Changing n to (n — 1), we get

$$\mu_{2n-2} = \frac{2^{n-1} . \sigma^{2n-2}}{\sqrt{\pi}} \Gamma\left(n - \frac{1}{2}\right)$$

$$\therefore \quad \frac{\mu_{2n}}{\mu_{2n-2}} = 2\sigma^2 . \frac{\Gamma\left(n + \frac{1}{2}\right)}{\left(n - \frac{1}{2}\right)} = 2\sigma^2\left(n - \frac{1}{2}\right) \qquad [\because \Gamma(r) = (r-1)\,\Gamma(r-1)]$$

$$\Rightarrow \qquad \mu_{2n} = \sigma^2 (2n-1)\, \mu_{2n-2}$$

This gives the recurrence relation for the moments of normal distribution

7.12 MEAN DEVIATION ABOUT MEAN:

Mean deviation about mean of normal distribution is given by

$$\text{M.D.(about mean)} = \int_{-\infty}^{\infty} |x-\mu| f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |x-\mu| e^{-(x-\mu)^2 / 2\sigma^2} dx$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| e^{-\frac{z^2}{2}} dz \qquad [putting \frac{x-\mu}{\sigma} = z]$$

$$= \frac{2\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| e^{-\frac{z^2}{2}} dz$$

Since the integrand $|z| e^{-\frac{z^2}{2}} dz$ is an even function of z .Also in [0, ∞], $|z|$ = z, so we have M.D.(about mean) =

$$\sqrt{\frac{2}{\pi}} \sigma \int_{0}^{\infty} z e^{-\frac{z^2}{2}} dz = \sqrt{\frac{2}{\pi}} \sigma \int_{0}^{\infty} e^{-t} dt \qquad [putting \frac{z^2}{2} = t]$$

$$= \sqrt{\frac{2}{\pi}} \sigma = \frac{4}{5} \sigma \quad \text{(approx)}$$

7.13 AREA PROPERTY (NORMAL PROBABILITY INTEGRAL):

If X~ N (μ, σ^2), then the probability that random value of X will lie between X = μ . and X = x_1 is given by:

$$P(\mu < X < x_1) = \int_{\mu}^{x_1} f(x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{x_1} e^{-(x-\mu)^2/2\sigma^2} dx$$

Put Z = $\frac{x-\mu}{\sigma}$ $\quad X - \mu = \sigma Z$

When $X = \mu$, $Z = 0$ and when $X = x_1$, $Z = \frac{x_1 - \mu}{\sigma} = z_1$, (say).

$$\therefore \quad P(\mu < X < x_1) = P(0 < Z < z_1) = \frac{1}{\sqrt{2\pi}} \int_0^{z_1} e^{-z^2/2}\, dz = \int_0^{z_1} \varphi(z)\, dz$$

Where $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}}$, is the probability density function of standard normal variate.

The definite integral $\int_0^{z_1} \varphi(z)$ is known as normal probability integral and gives the area under standard normal variate between the ordinates Z=0. and Z= z_1

In particular

$$P(\mu - \sigma < X < \mu + \sigma) = \int_{\mu-\sigma}^{\mu+\sigma} f(x)dx \quad \Rightarrow P(-1 < Z < 1) = \int_{-1}^{1} \varphi(z)dz$$

$$= 2\int_0^1 \varphi(z)dz = 2 \times 0.3413 = .6826$$

Similarly

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = \int_{\mu-2\sigma}^{\mu+2\sigma} f(x)dx \quad \Rightarrow P(-2 < Z < 2) = \int_{-2}^{2} \varphi(z)dz$$

$$= 2\int_0^2 \varphi(z)dz = 2 \times 0.4772 = 0.9544$$

and

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = \int_{\mu-3\sigma}^{\mu+3\sigma} f(x)dx \quad \Rightarrow P(-3 < Z < 3) = \int_{-3}^{3} \varphi(z)dz$$

$$= 2\int_0^3 \varphi(z)dz = 2 \times 0.49865 = 0.9973$$

7 .14 IMPORTANCE OF NORMAL DISTRIBUTION:

Normal distribution plays a very important role in statistical theory because of the following reasons

(i) Most of the distributions occurring in practice, e.g., Binomial, Poisson, Hyper geometric distributions, etc., can be approximated by normal distribution .

Moreover, many of the sampling distributions, e.g., Student's t, Snedecor's F, Chi-square distributions, etc., tend to normality for large samples.

(ii) Even if a variable is not normally distributed, it can sometimes be brought to normal form by simple transformation of variable.

(iii) The entire theory of small sample tests, viz., t, F, χ^2 tests, etc., is based on the fundamental assumption that the parent populations from which the samples have been drawn follow normal distribution.

(iv) Many of the distributions of sample statistics (e.g., the distributions of sample mean, sample variance, etc.) tend to normality for large samples and as such they can best be studied with the help of the normal curves.

(vi) Normal distribution finds large applications in Statistical Quality Control in industry for setting control limits.

(v) If $X \sim N(\mu, \sigma^2)$, then $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = P(-3 < Z < 3) = 0{\cdot}9973$

$\therefore$ $P(|Z| > 3) = 1 - P(|Z| \leq 3) = 0{\cdot}0027$ This property of the normal distribution forms the basis of entire Large Sample.

7.15 SUMMARY:

In previous lesson normal distribution was introduced .In this lesson we studied some important features of normal distribution like its mean, variance, median, m.g.f., recurrence relation for moments, linear combination of normal variates and area property which is the basis for the application of normal distribution.

7.16 SELF ASSESSMENTS:

1) Find mean deviation about mean of SNV.

2) If $X_1\ and\ X_2$ are SNV find the distribution of $X_1 + X_2$ and $X_1 - X_2$.

7.17 FURTHER READINGS:

1. Fundamentals of Mathematical Statistics : S C Gupta and V K Kapoor
2. New Mathematical Statistics : Bansi Lal and Sanjay Arora
3. An Introduction to Theory of Probability and Mathematical Statistics : V K Rohatgi
4. Continuous Univariate Distributions,N.K.Johnson and s. Kotz
5. Introduction to Mathematical Statistics R. V. Hogg and A. T. Craig.

****

UNIT-III LESSON-8

GAMMA DISTRIBUTION

STRUCTURE

8.1 OBJECTIVES:

The main objectives of this lesson are:

1. To introduce students to the concept of gamma distribution

2. To discuss some important features and properties of gamma distribution

8.2 INTRODUCTION:

In probability theory and statistics, the **gamma distribution** is a two-parameter family of continuous probability distributions. It has a scale parameter *a* and shape parameter λ. If λ is an integer, then the distribution represents an Erlang distribution. The gamma distribution is frequently a probability model for waiting times; for instance, in life testing, the waiting time until death is a random variable that is frequently modeled with a gamma distribution.

8.3 GAMMA DISTRIBUTION

Definition: A random variable X is said to have gamma distribution with parameter $\lambda > 0$, if its p.d.f. is given by :

$$f(x)=\begin{cases}\dfrac{e^{-x}x^{\lambda-1}}{\Gamma\lambda};\lambda>0,0<x<\infty\\ 0,otherwise\end{cases}$$

Remark :

A continuous random variable X have gamma distribution with two parameters a and λ if its p.d.f. is

$$f(x)=\begin{cases}\dfrac{a^{\lambda}e^{-ax}x^{\lambda-1}}{\Gamma\lambda};a>0,\lambda>0,0<x<\infty\\ 0,otherwise\end{cases}$$

8.4 M.G.F. OF GAMMA DISTRIBUTION :

By definition of M.G.F., we have

$$M_X(t)=E(e^{tx})=\int_0^\infty e^{tx}f(x)dx=\int_0^\infty e^{tx}\frac{e^{-x}x^{\lambda-1}}{\prod\lambda}$$

$$= \frac{1}{\Gamma\lambda}\int_0^\infty e^{tx}e^{-x}x^{\lambda-1}dx = \frac{1}{\Gamma\lambda}\int_0^\infty e^{-(1-t)x}x^{\lambda-1}dx$$

$$= \frac{1}{\Gamma\lambda}\cdot\frac{\Gamma\lambda}{(1-t)^\lambda}, |t| < 1$$

$$\therefore M_x(t) = (1-t)^{-\lambda}$$

8.5 CONSTANTS OF GAMMA DISTRIBUTION :

$$\mu_r^{'} = E(x) = \int_0^\infty e^r f(x)\, dx$$

$$= \int_0^\infty x^r \frac{e^{-x}x^{\lambda-1}}{\Gamma\lambda}dx$$

$$= \frac{1}{\Gamma\lambda}\int_0^\infty x^r e^{-x}x^{\lambda-1}dx = \frac{1}{\Gamma\lambda}\int_0^\infty x^{r+\lambda-1}e^{-x}dx$$

$$= \frac{1}{\Gamma\lambda}\cdot\Gamma(r+\lambda) \qquad \ldots\ldots(1)$$

Put $r = 1,2,...$ we have

$$\mu_1^{'} = \frac{\Gamma(\lambda+1)}{\lambda} = \lambda$$

$$\mu_2^{'} = \frac{\Gamma(\lambda+2)}{\lambda} = \lambda(\lambda+1)$$

$$\mu_3^{'} = \frac{\Gamma(\lambda+3)}{\lambda} = \lambda(\lambda+1)(\lambda+2)$$

$$\mu_4^{'} = \frac{\Gamma(\lambda+4)}{\lambda} = \lambda(\lambda+1)(\lambda+2)(\lambda+3)$$

$$\therefore \mu_2 = \mu_2^{'} - (\mu_1^{'})^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Thus for one parametric gamma distribution mean and variance are same.

Similarly $\mu_3 = 2\lambda, \mu_4 = 6\lambda + 3\lambda^2$

Hence $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{4\lambda^2}{\lambda^3} = \dfrac{4}{\lambda}$ and $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 3 + \dfrac{6}{\lambda}$.

8.6 TWO PARAMETER GAMMA DISTRIBUTION:

If X is two parametric gamma variate then its p.d.f. is given by

$$f(x,\lambda,a) = \begin{cases} \dfrac{a^{\lambda} e^{-ax} x^{\lambda-1}}{\Gamma\lambda}; a > 0, \lambda > 0, 0 < x < \infty \\ 0, \ otherwise \end{cases}$$

Here λ is the shape parameter and a is the scale parameter of gamma distribution.

The cumulative distribution function of gamma varite is given by

$$F(x,\lambda,a) = \int_0^x f(u,\lambda,a)du = \frac{\gamma\left(\lambda, x/_a\right)}{\Gamma\lambda}$$

Where $\gamma\left(\lambda, x/_a\right)$ ix lower incomplete gamma function.

8.7 MOMENTS OF TWO PARAMETER GAMMA DISTRIBUTION:

Then rth moment about origin μ_r' is given by

$$\mu_r' = E(X^r) = \int_0^\infty x^r f(x)dx$$

$$= \int_0^\infty x^r \ \frac{a^{\lambda} e^{-ax} x^{\lambda-1}}{\Gamma\lambda} \text{dx} \quad = \frac{a^{\lambda}}{\Gamma\lambda}\int_0^\infty x^{r+\lambda-1} e^{-ax} dx$$

$$= \frac{a^{\lambda}}{\Gamma\lambda} \ \frac{\Gamma(\lambda + r)}{a^{r+\lambda}} = \frac{\Gamma(\lambda + r)}{\Gamma\lambda \ a^r}$$

Mean $\mu_1' = \dfrac{\Gamma(\lambda+1)}{\Gamma\lambda a} = \dfrac{\lambda\Gamma\lambda}{a\Gamma\lambda} = \dfrac{\lambda}{a}$

$\mu_2' = \frac{\Gamma(\lambda+2)}{a^2\Gamma\lambda} = \frac{(\lambda+1)\lambda\Gamma\lambda)}{a^2\Gamma\lambda} = \frac{\lambda(\lambda+1)}{a^2}$

Var(X) = $\mu_2' - \mu_1'^2 = \frac{\lambda(\lambda+1)}{a^2} - \frac{\lambda^2}{a^2} = \frac{\lambda}{a^2}$

Obviously, Variance > Mean if a < 1

Variance = Mean if a = 1

Variance < Mean if a > 1

8.8 MODE OF GAMMA DISTRIBUTION:

For a continuous distribution with p.d.f. f(x) the mode is the value of the variate obtained by putting $f'(x) = 0$ provided $f''(x) < 0$ at the solution obtained by $f'(x) = 0$.

Since $f(x) = \frac{a^\lambda e^{-ax} x^{\lambda-1}}{\Gamma\lambda}$

Now, $f'(x) = 0$

$$\Rightarrow \frac{a^\lambda}{\Gamma(\lambda)}\left[(\lambda-1)x^{\lambda-2}e^{-ax} - ae^{-ax}x^{\lambda-1}\right] = 0$$

$$\Rightarrow x^{\lambda-2}e^{-ax}[\lambda - 1 - ax] = 0$$

$$\Rightarrow x = \frac{\lambda-1}{a}.$$

It can easily be seen that at $x = \frac{\lambda-1}{a}$, $f''(x) < 0$. Hence mode of the gamma distribution is $\frac{\lambda-1}{a}$.

8.9 LIMITING FORM OF GAMMA DISTRIBUTION:

Let X be gamma variate with parameter λ, then

$$E(X)=\lambda$$

And $Var(X)=\lambda$

The standard gamma variate is given by

$$Z=\frac{X-E(X)}{\sqrt{Var(X)}}=\frac{X-\lambda}{\sqrt{\lambda}}$$

Consider the M.G.F. of the variable Z , we have

$$M_{Z(t)}=E(e^{tz})=e^{-\frac{t\lambda}{\sqrt{\lambda}}}E(e^{xt/\sqrt{\lambda}})$$

$$=e^{-\frac{t\lambda}{\sqrt{\lambda}}}M_x\left(\frac{t}{\sqrt{\lambda}}\right)$$

$$=e^{-\frac{t\lambda}{\sqrt{\lambda}}}\left(1-\frac{t}{\sqrt{\lambda}}\right)^{-\lambda}$$

Taking log both sides we have

$$\log M_Z(t)=-\frac{t\lambda}{\sqrt{\lambda}}-\lambda\log\left(1-\frac{t}{\sqrt{\lambda}}\right)=-t\sqrt{\lambda}-\lambda\left(\frac{t}{\sqrt{\lambda}}+\frac{t^2}{2\lambda}+\ldots\right)$$

$$=-t\sqrt{\lambda}+t\sqrt{\lambda}+\frac{t^2}{2}+0(\lambda^{-1/2}),$$

Where $0(\lambda^{-1/2})$ are the terms containing $\lambda^{1/2}$ and higher powers of λ in the denominator.

Therefore

$$\lim_{\lambda\to\infty}\log M_z(t)=\frac{t^2}{2}$$

$\Rightarrow \lim_{\lambda \to \infty} M_z(t) = e^{\frac{t^2}{2}}$, which is the m.g.f. of a standard normal variate. Hence by uniqueness theorem of m.g.f., as $\lambda \to \infty$, standard gamma variate tends to standard normal variate. Hence Gamma distribution tends to Normal distribution for larger λ.

8.10 ADDITIVE PROPERTY OF GAMMA DISTRIBUTION :

If $X_1, X_2, ..., X_n$ are independent gamma variates with parameters $\lambda_1, \lambda_2, ..., \lambda_n$ respectively then $\sum_{i=1}^{n} X_i$ is also a Gamma variate with parameter $\sum_{i=1}^{n} \lambda_i$.

Proof:

Since X_i is gamma variate with parameter $\lambda_i; i = 1,2,...n$

Therefore, $M_{X_i}(t) = (1-t)^{-\lambda}$

Now Consider

$M_{\Sigma X_i}(t) = E\left(e^{t(X_1+X_2+...+X_n)}\right) = M_{X_1}(t) M_{X_2}(t)...M_{X_n}(t)$, since X_i's are independent

$= (1-t)^{-\lambda_1}(1-t)^{-\lambda_2}....(1-t)^{-\lambda_n} = (1-t)^{-\sum_{i=1}^{n}\lambda_i}$, which is the m.g.f. of gamma distribution with parameter $\sum_{i=1}^{n} \lambda_i$. Hence by uniqueness theorem of m.g.f., $\sum_{i=1}^{n} X_i$ is also a Gamma variate with parameter $\sum_{i=1}^{n} \lambda_i$.

8.11 APPLICATION OF GAMMA DISTRIBUTION

The gamma distribution has been used to model the size of insurance claims and rainfalls. This means aggregate insurance claims and the amount of rainfall accumulated in a reservoir are modeled by a gamma process. The gamma distribution is also used to model errors in multi-level Poisson regression models, because the combination of the Poisson distribution and a gamma distribution is a negative binomial distribution. It can be used for Internet traffic modeling.

1. The gamma is a flexible life distribution model that may offer a good fit to some sets of failure data. It is not, however, widely used as a life distribution model for common failure mechanisms.
2. The gamma does arise naturally as the time-to-first fail distribution for a system with standby exponentially distributed backups. If there are *n*-1 standby backup units and the system and all backups have exponential lifetimes with parameter λ , then the total lifetime has a gamma distribution .
3. A common use of the gamma model occurs in Bayesian reliability applications. When a system follows an HPP (exponential) model with a constant repair rate λ , and it is desired to make use of prior information about possible values of λ , a gamma Bayesian prior for λ is a convenient and popular choice.

8.12 ILLUSTRATIONS:

(1) Consumer demand for milk in a certain locality, per month, is known to be a gamma variate. If the average demand is 'α' liters and the most likely demand is 'β' liters ($\beta<\alpha$), what is the variance of the demand ?

Sol: If X~gam$(\lambda;a)$, then E(X)=$\frac{\lambda}{a}$, Var(X)=$\frac{\lambda}{a^2}$ and Mode= $\frac{\lambda-1}{a}$

Hence $\alpha=\frac{\lambda}{a}$ and $\beta=\frac{\lambda-1}{a}$

Solving these relations we get,

$$\frac{\lambda-1}{\lambda}=\frac{\beta}{\alpha}\Rightarrow 1-\frac{1}{\lambda}=\frac{\beta}{\alpha}\Rightarrow\lambda=\frac{\alpha}{\alpha-\beta}$$

Also $a=\dfrac{\lambda}{\alpha}=\dfrac{1}{\alpha-\beta}$

But Var(X) $=\frac{\lambda}{a^2}=\frac{\alpha}{\alpha-\beta}\cdot(\alpha-\beta)^2=\alpha(\alpha-\beta)$

(2) If $X\sim N(\mu,\sigma^2)$, obtain the p.d.f of $U=\dfrac{1}{2}\left(\dfrac{X-\mu}{\sigma}\right)^2$

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

$\therefore Z=\dfrac{x-\mu}{\sigma}\sim N(0,1)$ with p.d.f

$$f(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right):-\infty<z<\infty\,.$$

Thus, the distribution function $G(\cdot)$ of $U=\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2$ is given by

$$G(u)=P\{U\le u)=P\left(\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\le u\right)$$

$$=P\left(Z^2\le 2u\right)=P\left(-\sqrt{2u}\le Z\le\sqrt{2u}\right)$$

$$=P\left(Z\le\sqrt{2u}\right)-P\left(Z\le-\sqrt{2u}\right)=F\left(\sqrt{2u}\right)-F\left(-\sqrt{2u}\right)$$

Where $F(\cdot)$ is the c.d.f. of standard normal variate Z

Differentiating w.r.t. u, the p.d.f. $g(\cdot)$ of U is given byn :

$$g(u)=f(\sqrt{2u})\cdot\frac{d}{du}(\sqrt{2u})-f(-\sqrt{2u})\cdot\frac{d}{du}(-\sqrt{2u})$$

$$=\frac{1}{\sqrt{2u}}\left[f(\sqrt{2u})-f(-\sqrt{2u})\right]$$

$$=\frac{1}{\sqrt{2u}}\cdot 2f(\sqrt{2u})$$, since $f(u)$ is even function.

$$=\sqrt{\frac{2}{u}}\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\cdot 2u\right)$$

$=\frac{1}{\Gamma(1/2)}\cdot e^{-u}u^{(1/2)-1};u\ge 0$, which is the p.d.f of gamma distribution with parameter $\frac{1}{2}$.

8.13 SUMMARY:

In this lesson we introduced students to one and two parameter gamma distribution. Important characteristics like mean, variance, mode ,mgf, limiting distribution, additive property etc. have been discussed at length. Application of gamma distribution in variety of problems is also illustrated.

8.14 SELF ASSESSMENT:

(1) Let X and Y have the joint p.d.f.

$$f(x,y)=\begin{cases}\dfrac{e^{-(x+y)}x^3y^4}{\Gamma4\Gamma5};x>0,y>0\\ 0\ , \quad otherwise\end{cases}$$

Find (i) p.d.f of $U=\dfrac{X}{X+Y}$ $\quad (ii)E(U)\ (iii)Var(U)$

(2) A random sample of size n is drawn from a population whose p.d.f is

$$f(x)=\{\lambda e^{-\lambda x},\ x>0,\lambda>0$$

If $\overline{X}$ is the mean of the sample ,show that $n\lambda\overline{X}\ is$ a gamma variate and prove that $E(\overline{X})=\dfrac{1}{\lambda}\ and \quad S.E(\overline{X})=\dfrac{1}{\lambda\sqrt{n}}$

8.15 FURTHER READINGS:

1. Fundamentals of Mathematical Statistics : S C Gupta and V K Kapoor
2. New Mathematical Statistics : Bansi Lal and Sanjay Arora
3. Continuous Univariate Distributions,N.K.Johnson and s. Kotz.
4. Introduction to the Theory of Statistics: A M Mood ,F A Graybill and D C Boes

----

UNIT-III **LESSON-9**

BETA DISTRIBUTION

STRUCTURE

9.1 OBJECTIVES:

The objectives of this chapter are

1. To introduce students to beta distributions
2. To discuss important properties of beta distributions

9.2 INTRODUCTION:

In probability theory and Statistics, the beta distribution of firs kind is a continuous probability distributions defined on the interval (0, 1) parameterized by two positive shape parameters, typically denoted by m and n. The domain of the beta distribution can be viewed as a probability, and in fact the beta distribution is often used to describe the distribution of an unknown probability value — typically, as the prior distribution over a probability parameter, such as the probability of success in a binomial distribution or Bernoulli distribution. In fact, the beta distribution is the conjugate prior of the binomial distribution and Bernoulli distribution.

9.3 BETA DISTRIBUTION OF FIRST KIND:

Definition: A random variable X is said to have a beta distribution of first kind with parameters *m* and *n* $(m > 0, n > 0)$ if its p.d.f. is given by :

$$f(x) = \begin{cases} \dfrac{1}{\beta(m,n)} x^{m-1}(1-x)^{n-1}; (m,n) > 0, 0 < x < 1 \\ 0, \text{otherwise} \end{cases}$$

Where $\beta(m,n)$ is the beta function.

Remark : If we take $m = 1, n = 1$, we have

$$f(x) = \frac{1}{\beta(1,1)} = 1; 0 < x < 1,$$

which is the p.d.f of uniform distribution on $[0,1]$

9.4 CONSTANTS OF BETA DISTRIBUTION OF FIRST KIND :

$$\mu_r^{'} = E(x^r) = \int_0^1 x^r f(x)\, dx$$

$$= \int_0^1 x^r \frac{x^{m-1}(1-x)^{n-1}}{\beta(m,n)} dx$$

$$= \frac{1}{\beta(m,n)} \int_0^1 x^r x^{m-1}(1-x)^{n-1} dx$$

$$= \frac{1}{\beta(m,n)}\int_0^1 x^{r+m-1}(1-x)^{n-1}dx$$

$$= \frac{1}{\beta(m,n)}\cdot\beta(m+r,n)$$

$$= \frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)}\cdot\frac{\Gamma(m+r)\Gamma(n)}{\Gamma(m+r+n)}$$

$$= \frac{\Gamma(m+n)\Gamma(m+r)}{\Gamma(m)\Gamma(m+r+n)}$$

In particular,

Mean $\mu_1^{'} = \frac{\Gamma(m+1)}{\Gamma(m+n+1)}\cdot\frac{\Gamma(m+n)}{\Gamma(m)}$

$$= \frac{m\Gamma(m)\Gamma(m+n)}{(m+n)\Gamma(m+n)\Gamma(m)} = \frac{m}{m+n}$$

$$\mu_2^{'} = \frac{\Gamma(m+2)}{\Gamma(m+n+2)}\cdot\frac{\Gamma(m+n)}{\Gamma(m)}$$

$$= \frac{(m+1)m\Gamma(m)\Gamma(m+n)}{(m+n+1)(m+n)\Gamma(m+n)\Gamma(m)} = \frac{m(m+1)}{(m+n)(m+n+1)}$$

Similarly

$$\mu_3^{'} = \frac{m(m+1)(m+2)}{(m+n)(m+n+1)(m+n+2)}$$

$$\mu_4^{'} = \frac{m(m+1)(m+2)(m+3)}{(m+n)(m+n+1)(m+n+2)(m+n+3)}$$

Hence

$$\mu_2 = \mu_2^{'} - (\mu_1^{'})^2 = \frac{m(m+1)}{(m+n)(m+n+1)} - \frac{m^2}{(m+n)^2}$$

$$= \frac{m}{(m+n)^2(m+n+1)}\left[(m+n)(m+1) - m(m+n+1)\right]$$

$$= \frac{mn}{(m+n)^2(m+n+1)}.$$

Similarly, we have

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = \frac{2mn(n-m)}{(m+n)^2(m+n+1)(m+n+2)}$$

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

$$= \frac{3mn\{mn(m+n-6) + 2(m+n)^2\}}{(m+n)^2(m+n+1)(m+n+2)(m+n+3)}$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{4(n-m)^2(m+n+1)}{mn(m+n+2)^2}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3(m+n+1)mn(m+n-6) + 2(m+n)^2}{mn(m+n+2)(m+n+3)}$$

9.5 HARMONIC MEAN OF BETA DISTRIBUTION OF FIRST KIND:

$$\frac{1}{H} = E\left(\frac{1}{x}\right) = \int_0^1 \frac{1}{x} f(x)dx = \int_0^1 \frac{1}{x}\frac{x^{m-1}(1-x)^{n-1}}{\beta(m,n)}dx$$

$$= \frac{1}{\beta(m,n)}\int_0^1 \frac{1}{x} x^{m-1}(1-x)^{n-1}dx$$

$$= \frac{1}{\beta(m,n)}\int_0^1 x^{m-2}(1-x)^{n-1}dx$$

$$= \frac{1}{\beta(m,n)}\cdot\beta(m-1,n)$$

$$=\frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)}\cdot\frac{\Gamma(m-1)\Gamma(n)}{\Gamma(m+n-1)}$$

$$=\frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)}\cdot\frac{\Gamma(m-1)\Gamma(n)}{\Gamma(m+n-1)}$$

$$=\frac{(m+n-1)\Gamma(m+n-1)\Gamma(m-1)}{(m-1)\Gamma(m-1)\Gamma(m+n-1)}=\frac{m+n-1}{m-1}$$

9.6 BETA DISTRIBUTION OF SECOND KIND :

Definition: A random variable X is said to have a beta distribution of second kind with parameters *m* and *n* $(m > 0, n > 0)$ if its p.d.f. is given by :

$$f(x)=\begin{cases}\frac{1}{\beta(m,n)}\frac{x^{m-1}}{(1+x)^{m+n}}; (m,n)>0, 0<x<\infty\\ 0, \text{otherwise}\end{cases}$$

Remark :

Beta distribution of second kind is transformed to beta distribution of first kind by the transformation:

$$1+x=\frac{1}{y}\Rightarrow y=\frac{1}{1+x} \qquad (21.6.1)$$

Thus, if $X \sim \beta_2(m,n)$ then Y as defined in (21.6.1) is a $\beta_1(m,n)$.

9.7 CONSTANTS OF BETA DISTRIBUTION OF SECOND KIND :

$$\mu_r^{'}=E(x^r)=\int_0^{\infty}x^r f(x)dx$$

$$=\int_0^{\infty}x^r\frac{1}{\beta(m,n)}\cdot\frac{x^{m-1}}{(1+x)^{m+n}}dx$$

$$=\frac{1}{\beta(m,n)}\int_0^{\infty}x^r\cdot\frac{x^{m-1}}{(1+x)^{m+n}}dx$$

$$= \frac{1}{\beta(m,n)} \int_0^\infty \frac{x^{m+r-1}}{(1+x)^{m+n}} dx$$

$$= \frac{1}{\beta(m,n)} \int_0^\infty \frac{x^{m+r-1}}{(1+x)^{m+r+n-r}} dx$$

$$= \frac{1}{\beta(m,n)} \cdot \beta(m+r, n-r)$$

$$= \frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)} \cdot \frac{\Gamma(m+r)\Gamma(n-r)}{\Gamma(m+n)}$$

$$= \frac{\Gamma(m+r)\Gamma(n-r)}{\Gamma(m)\Gamma(n)}; n > r$$

In particular

$$\mu_1^{'} = \frac{\Gamma(m+1)\Gamma(n-1)}{\Gamma(m)\Gamma(n)} = \frac{m\Gamma(m)\Gamma(n-1)}{\Gamma(m)(n-1)\Gamma(n-1)} = \frac{m}{n-1}; n > 1$$

$$\mu_2^{'} = \frac{\Gamma(m+2)\Gamma(n-2)}{\Gamma(m)\Gamma(n)}$$

$$= \frac{(m+1)m\Gamma(m)\Gamma(n-1)}{\Gamma(m)(n-2)(n-1)\Gamma(n-1)}$$

$$= \frac{m(m+1)}{(n-1)(n-2)}; n > 2$$

Hence,

$$\mu_2 = \mu_2^{'} - (\mu_1^{'})^2 = \frac{m(m+1)}{(n-1)(n-2)} - \frac{m^2}{(n-1)^2}$$

$$= \frac{m}{n-1}\left[\frac{(n-1)(m+1) - m(n-2)}{(n-1)(n-2)}\right] = \frac{m(m+n-1)}{(n-1)^2(n-2)}$$

9.8 HARMONIC MEAN OF BETA DISTRIBUTION OF SECOND KIND :

$$\frac{1}{H} = E\left(\frac{1}{x}\right) = \int_0^\infty \frac{1}{x} f(x)dx = \int_0^\infty \frac{1}{x}\cdot\frac{1}{\beta(m,n)}\frac{x^{m-1}}{(1+x)^{m+n}}dx$$

$$= \frac{1}{\beta(m,n)}\int_0^\infty \frac{1}{x}\cdot\frac{x^{m-1}}{(1+x)^{m+n}}dx$$

$$= \frac{1}{\beta(m,n)}\int_0^\infty \frac{x^{m-2}}{(1+x)^{m+n}}dx$$

$$= \frac{1}{\beta(m,n)}\cdot\beta(m-1,n+1)$$

$$= \frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)}\cdot\frac{\Gamma(m-1)\Gamma(n+1)}{\Gamma(m+n)}$$

$$= \frac{\Gamma(m-1)\Gamma(n+1)}{\Gamma(m)\Gamma(n)}$$

$$= \frac{\Gamma(m-1)(n)\Gamma(n)}{(m-1)\Gamma(m-1)\Gamma(n)} = \frac{n}{m-1}$$

9.9 ILLUSTRATIONS :

(1): Let X and Y are independent gamma variates with parameters m and n respectively, then show that $U = X + Y$, $V = \frac{X}{Y}$ are independent and that U is a gamma variate with parameter $m + n$ and V is a Beta variate of second kind with parameters m and n.

Sol: Since X and Y are independent gamma variates with parameters m and n respectively, so the joint p.d.f. is

$$dF(x,y) = \frac{1}{\Gamma(m)\Gamma(n)}\cdot e^{-(x+y)}x^{m-1}y^{n-1}dxdy : 0 < (x,y) < \infty$$

Let $u = x + y, v = \frac{x}{y}$

$\Rightarrow y = \frac{u}{1+v}$ and $x = \frac{uv}{1+v} = u\left(1 - \frac{1}{1+v}\right)$

$\therefore J = \frac{\partial(x, y)}{\partial(u, v)} = -\frac{u}{(1+v)^2}$

Hence the joint p.d.f of U and V is :

$$dG(u, v) = \frac{1}{\Gamma(m)\Gamma(n)} \cdot e^{-u}\left(\frac{uv}{1+v}\right)^{m-1}\left(\frac{u}{1+v}\right)^{n-1}|J|dudv$$

$$= \frac{e^{-u}u^{m+n-1}}{\Gamma(m+n)} \cdot \frac{v^{m-1}}{\beta(m, n)(1+v)^{m+n}}$$

$$= dG_1(u) \cdot dG_2(v); 0 < u < \infty, 0 < v < \infty$$

This result shows that *U* and *V* are independent, *U* being a gamma vitiate with parameter $m + n$ and V is a Beta variate of second kind with parameters *m* and *n*.

(2): Let *X* and *Y* are independent gamma variates with parameters *m* and *n* respectively, then show that $U = X + Y$, $V = \frac{X}{X+Y}$ are independent and that *U* is a gamma vitiate with parameter $m + n$ and V is a Beta variate of first kind with parameters m and n.

Sol: since $X \sim \gamma(m)$ and $Y \sim \gamma(n)$. Therefore we have,

$$f_1(x)dx = \frac{1}{\Gamma(m)}e^{-x}x^{m-1}dx; 0 < x < \infty, m > 0$$

And $f_2(y)dy = \frac{1}{\Gamma(n)}e^{-y}y^{n-1}dx; 0 < y < \infty, n > 0$

Since *X* and *Y* are independent, therefore their joint p.d.f is given by:

$$dF(x, y) = \frac{1}{\Gamma(m)\Gamma(n)} \cdot e^{-(x+y)}x^{m-1}y^{n-1}dxdy : 0 < (x, y) < \infty$$

Let $u = x + y, v = \frac{x}{x+y}$

$\Rightarrow x = uv$ and $y = u(1-v)$

Therefore the jacobian of transformations is given by

$$J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & 1-v \\ u & -u \end{vmatrix} = -u$$

And u ranges from 0 to ∞ and v from 0 to 1, as x and y ranges from 0 to ∞. Hence the joint p.d.f of U and V is given by :

$$g(u,v) = \frac{1}{\Gamma(m)\Gamma(n)} \cdot e^{-u} \left(\frac{uv}{1+v}\right)^{m-1} \left(\frac{u}{1+v}\right)^{n-1} |J| dudv$$

$$= \frac{1}{\Gamma(m)\Gamma(n)} \cdot e^{-u} u^{m+n-1} v^{m-1} (1-v)^{n-1} dudv$$

$$= \frac{e^{-u} u^{m+n-1}}{\Gamma(m+n)} \cdot \frac{v^{m-1}(1-v)^{n-1}}{\beta(m,n)}$$

$$= dG_1(u) \cdot dG_2(v); 0 < u < \infty, 0 < v < 1$$

Thus we conclude that U and V are independently distributed, U as a $\gamma(m+n)$ and V as a $\beta_1(m,n)$.

(3): If $X \sim \beta_1(m,n)$ and $Y \sim \gamma(a, m+n)$ be independent random variables then Show that $XY \sim \gamma(a,m)$.

Sol: We are given that $X \sim \beta_1(m,n)$ and $Y \sim \gamma(a, m+n)$ are independent random variables. So their joint p.d.f. is :

$$f(x,y) = \frac{1}{B(m,n)} \cdot x^{m-1} (1-x)^{n-1} . \frac{a^{m+n}}{\Gamma(m+n)} e^{-ay} y^{m+n-1}; \; 0 < x < 1, \; 0 < y < \infty$$

Let $u = xy$ and $v = x$ so that $x = v$ and $y = \frac{u}{v}$

The Jacobian of transformations is given by

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ \frac{1}{v} & \frac{-u}{v^2} \end{vmatrix} = -\frac{1}{v}$$

Thus the joint p.d.f of *U* and *V* is given by :

$$g(u, v) = \frac{1}{B(m, n)} v^{m-1} (1-v)^{n-1} \cdot \frac{a^{m+n}}{\Gamma(m+n)} e^{-a(u/v)} \left(\frac{u}{v}\right)^{m+n-1} |J| \ ; 0 < u < \infty, 0 < v < 1$$

$$= \frac{a^{m+n}}{B(m,n)\Gamma(m+n)} \cdot v^{m-1}(1-v)^{n-1} \cdot e^{-a(u/v)} \left(\frac{u}{v}\right)^{m+n-1} \frac{1}{v}$$

$$= \frac{a^{m+n}\Gamma(m+n)}{\Gamma(m)\Gamma(n)\Gamma(m+n)} \cdot v^{-n-1}(1-v)^{n-1} \cdot e^{-a(u/v)} (u)^{m+n-1}$$

$$\frac{a^{m+n}}{\Gamma(m)\Gamma(n)} \cdot v^{-n-1}(1-v)^{n-1} \cdot e^{-a(u/v)} (u)^{m+n-1}$$

Integrating w.r.t, '*v*' the marginal p.d.f. of *U* is :

$$g_1(u) = \frac{a^{m+n}(u)^{m+n-1}}{\Gamma(m)\Gamma(n)} \int_0^1 \frac{(1-v)^{n-1} \cdot e^{-q(u/v)}}{v^{n+1}} dv$$

$$= \frac{a^{m+n}(u)^{m+n-1}}{\Gamma(m)\Gamma(n)} \int_0^1 \frac{1}{v^2} \left(\frac{1}{v} - 1\right)^{n-1} e^{-a(u/v)} dv$$

Put $\frac{1}{v} - 1 = t$ so that $dv = -\frac{1}{(1+t)^2} dt$

$$\therefore g_1(u) = \frac{a^{m+n} u^{m+n-1} e^{-au}}{\Gamma(m)\Gamma(n)} \int_0^\infty t^{n-1} e^{-aut} dt$$

$$= \frac{a^{m+n}u^{m+n-1}e^{-au}}{\Gamma(m)\Gamma(n)}\cdot\frac{\Gamma(n)}{(au)^{n}} = \frac{a^{m}}{\Gamma(m)}\cdot u^{m-1}e^{-au}$$ $;0 < u < \infty$, which is the p.d.f. of gamma distribution with parameters *a* and *m*.

Hence $XY \sim \gamma(a,m)$

(4): Given the Incomplete beta function

$$B_x(l,m) = \int_0^x x^{l-1}(1-x)^{m-1}dx$$

and $I_x(l,m) = {B_x(l,m)}\big/{B(l,m)}$,Show that

$$I_x(l,m) = 1 - I_{1-x}(m,l)$$

Sol. We have

$$I_x(l,m)\ B(l,m) = B_x(l,m) = \int_0^x x^{l-1}(1-x)^{m-1}dx \qquad (A)$$

$$= \int_0^1 x^{l-1}(1-x)^{m-1}dx - \int_x^1 x^{l-1}(1-x)^{m-1}dx$$

$$= B(l,m) - \int_x^1 x^{l-1}(1-x)^{m-1}dx$$

Put $1 - x = y$, then

$$I_x(l,m)\ B(l,m) = B(l,m) - \int_{1-x}^0 (1-y)^{l-1}y^{m-1}(-dy)$$

$$= B(l,m) - \int_0^{1-x}(1-y)^{l-1}y^{m-1}dy$$

$$= B(l,m) - B_{1-x}(m,l) = B(l,m) - \quad 1 - I_{1-x}(m,l)\ B(l,m)$$

Since $\mathrm{B(l,m) = B(m,l)}$, on dividing both sides by $\mathrm{B(l,m)}$,we get

$$\mathrm{I_x(l,m) \ = \ 1 - I_{1-x}(m,l)}$$

9.10 SUMMARY:

In this section we introduced two beta distributions, beta distribution of firs and second kind. Obtained some characteristics of these distributions and established their relationships with some other distributions. Illustrations are also given at the end of the section.

10.11 SELF ASSESSMENTS:

1. If $X \sim gamma(\lambda, \mu)$ and $Y \sim gamm\text{a}\,(\lambda, \nu)$ are independent random variables, show that $\dfrac{X}{Y} \sim \beta_2(\mu, \nu)$.

2. If $X \sim gamm\text{a}\,(\lambda, a)$ a*nd* $Y \sim gamm\text{a}\,(\lambda, b)$ are independent random variables, show that

$$U = X + Y,\, V = \frac{X}{X+Y} \quad are \;\; independen \; t, \mathrm{U} \;\; is \;\; gamma \;\; (\lambda, a+b)\, and \;\; V \;\; is \beta_1(a,b)$$

3. Let X be a beta variate with E(X) = 1/4 and Var(X) =1/8 ,find the values of the parameters of X.

9.12 FURTHER READINGS:

1. Fundamentals of Mathematical Statistics : S C Gupta and V K Kapoor
2. New Mathematical Statistics : Bansi Lal and Sanjay Arora
3. Continuous Univariate Distributions,N.K.Johnson and s. Kotz.
4. Introduction to the Theory of Statistics: A M Mood ,F A Graybill and D C Boes

****

UNIT-III LESSON - 10

EXPONENTIAL DISTRIBUTION AND ITS PROPERTIES

STRUCTURE

10.1 OBJECTIVES:

The objectives of this lesson are

1. To make students aware of one of the most important distribution in Statistics

2. To throw light on important properties of exponential distribution.

3. To discuss the applicability of exponential distribution in variety of situations.

10.2 INTRODUCTION:

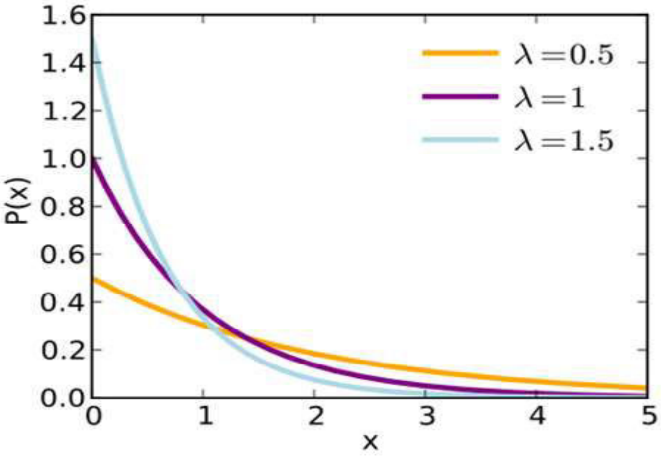
Exponential distribution is next important distribution in Statistics after the normal distribution. It can be obtained from gamma distribution as a particular case. Exponential distribution is the most widely used distribution in life testing experiments due to its Memoryless property, which states that if the life time of a component is exponentially distributed and the component has been used for some specified time ,then the distribution of the residual life time has the same exponential distribution. i.e. an old functioning component is as good as a new functioning component.

10.3 PROBABILITY DENSITY FUNCTION:

A continuous random variable X is said to have exponential distribution if its probability density function **(pdf)** is given by

$$f(x;\lambda) = \begin{matrix} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{matrix}$$

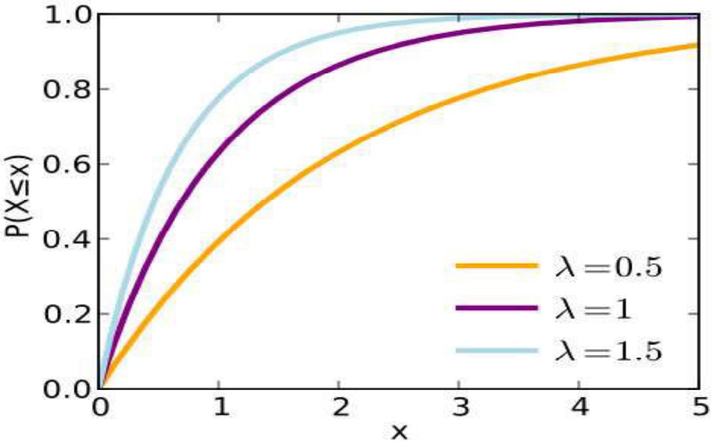
Here $\lambda > 0$ is the parameter of the distribution, often called the *rate parameter*. The distribution is supported on the interval [0, ∞). If a random variable *X* has exponential distribution, we write $X \sim \text{Exp}(\lambda)$.



10.4 CUMULATIVE DISTRIBUTION FUNCTION :

The cumulative distribution function is given by

$$f(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$



10.5 ALTERNATIVE PARAMETERIZATION:

A commonly used alternative parameterization is to define the probability density function (pdf) of an exponential distribution as

$$f(x;\beta) = \begin{array}{ll} \frac{1}{\beta} e^{-x/\beta} & x \geq 0 \\ 0 & x < 0 \end{array}$$

where b > 0 is a scale parameter of the distribution and is the reciprocal of the rate parameter, λ, defined above. In this specification, b is a survival parameter in the sense that if a random variable X is the duration of time that a given biological or mechanical system manages to survive and X ~ Exponential(b) then E[X] = b . That is to say, the expected duration of survival of the system is b units of time. The parameterisation involving the "rate" parameter arises in the context of events arriving at a rate , when the time between events (which might be modelled using an exponential distribution) has a mean of $\beta = \lambda^{-1}$.

10.6 OCCURRENCE AND APPLICATIONS:

The exponential distribution occurs naturally when describing the lengths of the inter-arrival times in a homogeneous Poisson process.

The exponential distribution may be viewed as a continuous counterpart of the geometric distribution, which describes the number of Bernoulli trials necessary for a *discrete* process to change state. In contrast, the exponential distribution describes the time for a continuous process to change state.

In real-world scenarios, the assumption of a constant rate (or probability per unit time) is rarely satisfied. For example, the rate of incoming phone calls differs according to the time of day. But if we focus on a time interval during which the rate is roughly constant, such as from 2 to 4 p.m. during work days, the exponential distribution can be used as a good approximate model for the time until the next phone call arrives. Similar caveats apply to the following examples which yield approximately exponentially distributed variables:

- The time until a radioactive particle decays, or the time between clicks of a geiger counter
- The time it takes before your next telephone call
- The time until default (on payment to company debt holders) in reduced form credit risk modeling

Exponential variables can also be used to model situations where certain events occur with a constant probability per unit length, such as the distance between mutations on a DNA strand, or between roadkills on a given road.

In queuing theory, the service times of agents in a system (e.g. how long it takes for a bank teller etc. to serve a customer) are often modeled as exponentially distributed variables. (The inter-arrival of customers for instance in a system is typically modeled by the Poisson distribution in most management science textbooks.) The length of a process that can be thought of as a sequence of several independent tasks is better modeled by a variable following the Erlang distribution (which is the distribution of the sum of several independent exponentially distributed variables).

10.7 MEAN, VARIANCE AND MEDIAN

The mean or expected value of an exponentially distributed random variable X with rate parameter λ is given by

$$E(X) = \lambda\int_0^\infty \; xe^{-\lambda\lambda} dx = \frac{1}{\lambda}$$

In light of the examples given above, this makes sense : if you receive phone calls at an average rate of 2 per hour, then you can expect to wait half an hour for every call.

The variance of X is given by

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

Now $E(X^2) = \lambda\int_0^\infty \; x^2 e^{-\lambda\lambda} dx = \frac{2}{\lambda^2}$

Hence Var(X) = $\text{E}(\text{X}^2) \text{ - } [\text{E}(\text{X}]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$

Let 'm' be the median of X. Then median of X is given by

$$\text{Med(X)} = \text{P(X<m)} = \frac{1}{2}$$

$$\text{i.e.}\lambda\int_0^\infty \; e^{-\lambda x} dx = \frac{1}{2}$$

$$(\text{i} - \text{e}^{-\lambda \text{m}}) = \frac{1}{2}$$

$$\Rightarrow \text{e}^{-\lambda \text{m}} = \frac{1}{2}$$

$$\Rightarrow \text{m} = \frac{\log^2}{\lambda} < \frac{1}{\lambda} = \text{E(X)}$$

Thus for exponential distribution median is less than mean.

10.8 MGF OF EXPONENTIAL DISTRIBUTION :

The m.g.f. of exponential distribution is given by

$$\text{M}_\text{x}(\text{t}) = \text{E}[\text{e}^{\text{tx}}] = \lambda\int_0^\infty \text{e}^{\text{tx}}\text{e}^{-\lambda \text{x}}\text{dx} = \frac{\lambda}{\lambda - \text{t}}, \lambda > \text{t}$$

The rth moment about origin is given by

$$\mu_r^{'} = E(x^r) = \text{coeffcicient of } \frac{t^r}{r!} \text{in } M_X(t) = \frac{r!}{\lambda^r}; = 1, 2, 3........$$

In particular mean

$$\mu_1^{'} = \frac{1}{\lambda} \text{and variance } \mu_2 = \mu_2^{'} - \mu_1^{'2} = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

10.9 SHIFTEN EXPONENTIAL DISTRIBUTION

A non negative continuous random variable X is said to have Shifted Exponential Distribution if its pdf is given by

$$f(x) = \lambda e^{-\lambda(x-\theta)}, \lambda > 0, x \geq \theta; f(x) = 0, \text{elsewhere}$$

The mgf in this case is given by

$$M_X(t) = E(e^{tX)} = \lambda \int_0^{\infty} e^{tx} e^{-(X-\theta)} dx \qquad e^{\lambda\theta} \left[1 - \frac{t}{\lambda}\right]^{-1}$$

$$= e^{\lambda\theta} \sum_{\lambda=0}^{\infty} \left(\frac{t}{\lambda}\right)^r \text{and } M_r^{'} = e^{\lambda\theta} . \frac{r!}{\lambda r}$$

Theorem 11.9.1 :

If $X_1, X_2, ...X_n$ are independent random variable having exponential distribution with parameter λ_i; i = 1, 2,.................n, then Z = min $(X_1, X_2...........X_n)$ has exponential distribution with parameter $\sum_{i=1}^{n} \lambda_i$

Proof : Let $G_z(z)$ denotes the distribution function of Z = min $(X_1, X_2...........X_n)$

Then $G_Z(z) = P(Z \leq z) = 1 - P\,(Z > z)$

$= 1 - P\,[\min(X_1, X_2...........X_n) > z]$

$= 1 - P[X_i > z; i=1, 2............n]$

$$1 - \prod_{i=1}^{n} P(X_i > z) = 1 - \prod_{i=1}^{n} [1 - P(X_i \leq z)]$$

$$= 1 - \prod_{i=1}^{n} [1 - F_{xi}\ (z)]$$

Where F(.) is the distribution function of X_i and is equal to $= 1 - e^{-\lambda_i z}$

So $G_z(z) = 1 - \prod_{i=1}^{n}[1-(1-e^{-\lambda_i z})]$

$= 1 - e^{\sum_{i=1}^{n}\lambda_i z}, z > 0$

Which is the distribution function of an exponential variate with parameter $\sum_{i=1}^{n}\lambda_i$

Hence $Z = \min(X_1, X_2..........X_n)$ has exponential distribution with parameter $\sum_{i=1}^{n}\lambda_i$

Remark : If $X_1, X_2..........X_n$ are i.i.d. exponential with parameter λ, then

$Z = \min(X_1, X_2..........X_n)$ has exponential distribution with parameter $n\lambda$.

11.10 RELATION WITH UNIFORM DISTRIBUTION :

If X~exp(λ), then Y = $\lambda^{-\lambda x}$ is U(0,1)

Proof: Given X~exp(λ) so pdf of X is given by

$$f(x) = \lambda e^{-\lambda\lambda}, \qquad \lambda > 0, x \geq 0$$

Since $y = e^{-\lambda\lambda}$

$$\left|\frac{dy}{dx}\right| = \lambda e^{-\lambda x}, \quad \left|\frac{dy}{dx}\right| = \frac{1}{\lambda}e^{\lambda x}$$

Since $0 \leq x < \infty, \quad \Rightarrow 0 \leq y < 1$

Now pdf of $y = e^{-\lambda n}$ is given by

$$f(y) = \left|\frac{dx}{dy}\right| f(x) = \frac{1}{\lambda}e^{\lambda x}\lambda e^{-\lambda x} = 1$$

Thus $f(y) = 1, \quad 0 \leq y < 1$

Hence $y = e^{-\lambda x}$ is $U(0,1)$

10.11 MEMORYLESS PROPERTY:

An important property of the exponential distribution is that it is memoryless. This means that if a random variable X is exponentially distributed, its conditional probability obeys

$$P\{X > s{+}t \mid X > t\} = P\ \{X > s\} \text{ for all s, t} \geq 0$$

If X denotes the life time of an component then memoryless property says that the probability that the component will survive for at least s+t time units given that it has survived t time units is the same as the initial probability that it survives for at least s time units. In other words if the component is working at time t, then the distribution of the remaining life time is same as the original life time distribution i.e. the component does not age or it does not remember the time for which it has been used for.

Proof: We are to prove that

$$P\{X > s{+}t \mid X > t\} = P\ \{X > s\} \text{ for all s, t} \geq 0 \qquad (1)$$

$$\text{Consider } P\{X > s{+}t \mid X > t\} = \frac{P\{x > s + t, X > t\}}{P\{X > t\}}$$

Since s, t $\geq$ 0 so $P\{X > s + t, X > t\} = P\{X > s + t\}$

$$\text{Therefore } P\{X > s{+}t \mid X > t\} = \frac{P\{X > s + t\}}{P\{X > t\}} \qquad (2)$$

$$\text{Now } P\ \{X > s + t\} = \lambda \int_{s+t}^{\infty} e^{-\lambda x} dx = e^{-\lambda(s+t)}$$

Similarly, $P\{X > t\} = e^{-\lambda x}$ and $P\{X > s\} = e^{-\lambda s}$

Substituting these values in (2), we get

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

This completes the proof.

The exponential distributions and the geometric distributions are the only probability distributions exhibiting memoreless property.

10.12 SUMMARY:

In this lesson we have introduced students with the various forms of exponential distribution along with some of its important properties, its relations with some other distributions. The situations for the occurrence and applicability of the exponential distribution have been discussed. Application of exponential distribution has been explained by taking examples.

10.13 ILLUSTRATION:

1) If X has exponential distribution with mean 2, find P[X<1|X<2]

Sol. If X $\sim$ exp (λ), then we have f(x) = $\lambda e^{-\lambda n}$, $\lambda > 0$

$$\text{Now P[X<1|X<2]} = \frac{P[x<1, X<2]}{P[X<2]} = \frac{P[x<1]}{P[X<2]}$$

$$= \frac{\int_0^1 \lambda e^{-\lambda x} dx}{\int_0^2 \lambda e^{-\lambda x} dx} = \frac{1-e^{-\lambda}}{1-e^{-\lambda}}$$

$$= \frac{(1-e^{-\lambda})}{(1-e^{-\lambda})(1+e^{-\lambda})} = \frac{1}{(1+e^{-\lambda})}$$

Now since the mean of the distribution $= \frac{1}{\lambda}$ is 2, so $\lambda = \frac{1}{2}$

Hence required probability = $= \frac{1}{(1+e^{-1/2})} = \frac{\sqrt{e}}{(1+\sqrt{e})}$

2) If X $\sim$ exp (λ), such that (PX $\leq$ 1) = P(X>1), then find Var(X)

Sol. Given (PX $\leq$ 1) = P(X > 1)

But P(X > 1) = 1 – P (X $\leq$ 1)

Hence P (X $\leq$ 1) = 1 – P (X $\leq$ 1)

$\Rightarrow$ 2 P (X $\leq$ 1) = 1 or P (X $\leq$ 1) = $\frac{1}{2}$

i.e. $\lambda\int_0^1 e^{-\lambda x}dx = \frac{1}{2}$

$(1 - e^{-\lambda}) = \frac{1}{2}$

$\Rightarrow e^{-\lambda} = \frac{1}{2}$

$\Rightarrow \lambda = \log 2$

But Var (X) = $= \frac{1}{\lambda^2} = \frac{1}{(\log 2)^2}$

10.14 SELF ASSIGNMENT:

(1) The length of the time (in minutes) that a lady speaks on the telephone is a random variable with p.d.f.

$$f(x) = Ae^{-\frac{x}{5}}, x > 0; f(x) = 0, \text{elsewhere}$$

(a) Evaluate A. What is the probability that the number of minutes she takes on phone is

(i) more than 10 minutes (ii) less than 5 minutes (iii) between 5 and 10 minutes.

(2) Suppose that the amount of time one spends in bank is exponentially distributed with mean 10 minutes. Find the probability that a customer will spend more than 15 minutes in bank.

(3) The time T required to repair a machine is an exponential variate with mean ½ hours.Find

(a) the probability that repair time exceeds ½ hour.

(b) the probability that repair time exceeds 12½ hour given that its duration exceeds 12 hours.

(3) If the pdf of X is given by

$$f(x)=\begin{cases} ce^{-2x}, 0 < x < \infty \\ 0, \text{otherwise} \end{cases}$$

Find c, and P(X>2)

10.15 FURTHER READINGS:

1. Continuous Univariate Distributions,N.K.Johnson and s. Kotz

2. Introduction to the Theory of Statistics: A M Mood ,F A Graybill and D C Boes

3. Fundamentals of Mathematical Statistics : S C Gupta and V K Kapoor

4. New Mathematical Statistics : Bansi Lal and Sanjay Arora

5. Introduction to Probability Models : Sheldon M Ross

******

UNIT-V **LESSON-11**

SOME MOMENTS INEQUALITIES

STRUCTURE

11.1 OBJECTIVES:

This chapter aims at

1. Introducing students to some important moments inequalities.
2. Familiarising them to the concept of convex function.
3. Making them capable to use these inequalities in various Statistical problems.

11.2 INTRODUCTION:

Moments inequalities play important role in theory of Statistics. A lot of literature on moments inequalities along with their importance and application is available. In the present section we shall deal with some of the moments inequalities like Chebyshev's inequality, Markov's inequality and Jensen's inequality and also discuss their significance and applications.

11.3 CHEBYSHEV'S INEQUALITY:

In probability theory, Chebyshev's inequality (also spelled as Tchebysheff's inequality) guarantees that in any data sample or probability distribution,"nearly all" values are close to the mean - the precise statement being that no more than $\frac{1}{k^2}$ of the distribution's values can be more than k times standard deviations away from the mean. The inequality is named after Russian mathematician Pafnuty Chebyshev, although it was first formulated by his friend and colleague Irénée-Jules BienayméThe inequality has great utility because it can be applied to completely arbitrary distributions (unknown except for mean and variance), for example it can be used to prove the weak law of large numbers.

Statement : If X is a random variable with mean μ and variance σ^2 then for any positive number k, we have

$$P\{|X-\mu|\geq k\sigma\}\leq\frac{1}{k^2}$$

or

$$P\{| X - \mu | < k\sigma\} \geq 1-\frac{1}{k^2}$$

“The probability that the outcome of an experiment with the random variable X will fall more than k standard deviations beyond the mean of X, μ, is less than $\frac{1}{k^2}$”

Or : “The proportion of the total area under pdf of X outside of k standard deviations from the mean μ is at most $\frac{1}{k^2}$”

Proof : The inequality is proved here assuming X to be a continuous random variable ,the inequality can also be proved by taking X to be a discrete random variable ,the only difference will be that integral sign is to be replaced by summation sign.

Let S be the sample space for a random variable, X , and let $f_X(x)$ stand for the pdf of x. . Let R_1, R_2 and R_3 partition S such that for every sample point x in S

$$x \in \begin{cases} R_1 \text{when } x < \mu - k\sigma \\ R_2 \text{when } |x-\mu| \le k\sigma \\ R_3 \text{when } x > \mu - k\sigma \end{cases}$$

Then we have

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

$$\sigma^2 = \int_{R_1} (X-\mu)^2 f(x)dx + \int_{R_2} (X-\mu)^2 f(x)dx + \int_{R_3} (X-\mu)^2 f(x)dx$$

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

$$\sigma^2 \ge \int_{-\infty}^{\mu-k\sigma} (X-\mu)^2 f(x)dx + \int_{\mu+k\sigma}^{\infty} (X-\mu)^2 f(x)dx \qquad (25.3.1)$$

Since $X \le \mu\text{-}k\sigma \; and \; X \ge \mu + k\sigma \Leftrightarrow | X - \mu | \ge k\sigma$ (25.3.2)

Using (25.3.2) in (25.3.1), we get

$$\sigma^2 \ge k^2\sigma^2 \left[\int_{-\infty}^{\mu-k\sigma} f(x)dx + \int_{\mu+k\sigma}^{\infty} f(x)dx \right]$$

or

$$\sigma^2 \ge k^2\sigma^2 [P(X \le \mu - k\sigma] + P[| X \ge \mu + k\sigma)]$$

i.e. $\sigma^2 \ge k^2\sigma^2 P\{|X-\mu| \ge k\sigma\}$

$$\Rightarrow P\{| X - \mu | \ge k\sigma\} \le \frac{1}{k^2}$$

Which complete the proof.

Also since

$P[| X - \mu | \ge k\sigma] + P[| X - \mu | < k\sigma] = 1$, so, we have

$$P[|X-\mu|<k\sigma] = 1 \ - P[|X-\mu| \geq k\sigma] \geq 1-\frac{1}{k^2}$$

11.4 SIGNIFICANCE OF CHEBYSHEV'S INEQUALITY:

The significance of Chebyshev's Inequality is that it enables us to find the bounds on probabilities without actually knowing the distribution of X the only condition is that mean and variance of the random variable are known . When actual distribution is known then the desired probability can be computed exactly . Chebyshev's Inequality is not only valid for absolutely continuous and discrete distributions, but also for variates which do not fall in these two categories, provided their variances exist.

Remark: In particular if we take k σ = c, then Chebyshev's Inequality reduces to

$$P\{|X-\mu| \geq c\} \leq \frac{\text{var}(x)}{c^2} \text{i.e.} \, P\{|X-\mu| < c\} \leq \frac{\sigma^2}{c^2}$$

These are convenient forms for numerical problems.

11.5 MARKOV'S INEQUALITY:

In probability theory, Markov's inequality gives an upper bound for the probability that a non-negative function of a random variable is greater than or equal to some positive constant. It is named after the Russian mathematician Andrey Markov, although it appeared earlier in the work of Pafnuty Chebyshev (Markov's teacher), and many sources, especially in analysis, refer to it as Chebychev's inequality or Bienaymé's inequality.

Markov's inequality (and other similar inequalities) relate probabilities to expectations, and provide (frequently) loose but still useful bounds for the cumulative distribution function of a random variable.

An example of an application of Markov's inequality is the fact that (assuming incomes are non-negative) no more than 1/5 of the population can have more than 5 times the average income.

Statement : Let $g(X) = |X|^r$ be a non negative function of a random variable X, then for any $\varepsilon > 0$,we have

$$P[|X| \geq \varepsilon] \leq \frac{E|X|^2}{\varepsilon^2}$$

Proof : Let us donate by the set $S = \{x : | X |^r \geq \varepsilon^r\}$ and S` as the compliment of S. Then, we have

$$P(X \in S) = \int_s dF(x) \text{ where F(x) is the distribution of X.}$$

Also $\quad E\{| X |^r\} \quad = \int_{-\infty}^{\infty} | x |^r \, dF(x)$

$$= \int_{S`} |x|^r dF(x) + \int_{S`} |x|^r dF(x)$$

$$\Rightarrow E\{|X|^r \geq \int_s |x|^r dF(x) = \int_{|X|^r \geq \varepsilon^r} |x|^r dF(x)$$

$$\Rightarrow E\{|X|^r \geq \varepsilon^r \int_{|X|^r \geq \varepsilon^r} |x|^r dF(x) = \varepsilon^r P[|X|^r \geq \varepsilon^r]$$

$$\Rightarrow E[|X|^r \geq \varepsilon^r] \leq \frac{E|X|^r}{\varepsilon^r}$$

Or $\quad \Rightarrow E[|X| \geq \varepsilon] \leq \dfrac{E|X|^r}{\varepsilon^r}$

11.6 CONVEX FUNCTION :

A continuous function f(x) on the interval I is said to be convex if for every x_1 and x_2, $\dfrac{x_1 + x_2}{2} \in I$, we have

$$f\left(\frac{x_1 + x_2}{2}\right) \leq \frac{1}{2} f(x_1) + \frac{1}{2} f(x_2)$$

If x_1 and $x_2 \in$ I, then $\dfrac{x_1 + x_2}{2} \in$ I.

A lternatively, A continuous f(x) on the interval I is said to be convex if for every x_1 and $x_2 \in$ I, we have

$$f(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2); \lambda_1 \geq 0, \lambda_2 \geq 0, \lambda_1 + \lambda_2 = 1$$

Remark : If f is twice differentiable i.e. if f``(x) exists and f`(x) $\geq$ 0 then f is convex.

11.7 JENSEN'S INEQUALITY:

In mathematics, Jensen's inequality, named after the Danish mathematician Johan Jensen, relates the value of a convex function of an integral to the integral of the convex function. It was proved by Jensen in 1906. Given its generality, the inequality appears in many forms depending on the context, some of which are presented below. In its simplest form the inequality states that the convex transformation of a mean is less than or equal to the mean after convex transformation; it is a simple corollary that the opposite is true of concave transformations.

Jensen's inequality generalizes the statement that the secant line of a convex function lies above the graph of the function, which is Jensen's inequality for two points: the secant line consists of weighted means of the convex function, $tf(x) + (1 - t)f(y)$, while the graph of the function is the convex function of the weighted means, $f(tx + (1 - t)y)$.

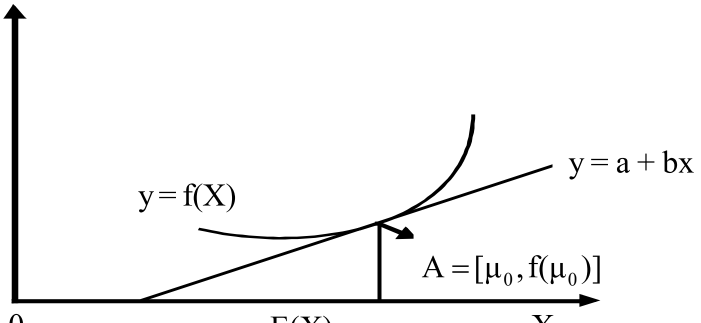
There are also converses of the Jensen's inequality, which estimate the upper bound of the integral of the convex function.

In the context of probability theory it is generally stated in the following form: if X is a random variable and is a convex function, then $E[f(X)] \geq f\,E(X)$

Statement: If f is continuous and convex function and X is a random variable with finite mean $\mu = E(X)$, then

$$E[f(X)] \geq f\,E(X)$$

Proof: Let $y = a + bx$ be a tangent to the curve $y = f(x)$ at the point $x_0 = E(X)$. Since f is continuous and convex function, the curve lies above the tangent so $f(x) \geq a + bx$, for all x. Therefore,



$f(X) \geq a + bX$

Hence $E[f(X)] \geq a + b\,E(X) = a + bx_0 = y_0 = f\;E(X)$

Thus $E[f(X)] \geq f\;E(X)$

Remaks :

(1) Jensen's inequality becomes equality if f is a linear function of x i.e. if

$P[f(X)] = a + bX] = 1$ for some a and b.

(2) If f is continuous and concave function and X is a random variable with finite mean $\mu = E(X)$, then

$E[f(X)] \leq f\;E(X)$

11.8 ILLUSTRATIONS:

(1) If X is the number scored in a throw of a fair die, show that the Chebyshev's inequality gives

$P[|\,X - \mu\,| > 2.5] < 0.47$

Where μ is the mean of X,while the actual probability is zero.

Sol. Here X is a random variable which takes the values 1,2,……6 each with probability 1/6.Hence

$$E(X) = \mu = \frac{1}{6}(1+2+3+.....+6) = \frac{7}{2}$$

$$E(X^2) = \frac{1}{6}(1^2 + 2^2 + ... + 6^2) = \frac{91}{6}$$

$$\text{Var}\,(X) = E(X^2) - [\,E(X)]^2 = \frac{91}{6} - \frac{49}{4} = 2.9167$$

Now for $k > 0$,Chebyshev's inequality gives

$$P[|\,X - \mu\,| > k] < \frac{\text{Var}(X)}{k^2}$$

Choosing $k = 2.5$, we get

$$P[|X-\mu|>2.5]<\frac{2.9167}{6.25}=0.47$$

The actual probability is given by

P = P[|X - 3.5| >2.5]

P[X lies outside the limits (3.5 - 2.5,3.5 + 2.5)]

i.e. P[X lies outside the limits(1,6)]

But since X is the number appearing on the dice when thrown ,so it can not be outside (1,6)

Hence required probability is zero.

(2) A symmetric dice is thrown 600 times. Find a lower bound for the probability of getting 80 to 120 sixes.

Sol. Let S be the total number of successes.

Then S ~ B (n, p) with n = 600, p = 1/6 so that q = 5/6

$$E(S)=npq=\frac{600}{6}x\frac{5}{6}=\frac{500}{6}$$

Using Chebyshev's inequality we get

$$P[|S-E(S)|\le k]\ge 1-\frac{Var(S)}{k^2}$$

$$\Rightarrow P[|S-100|\le k]\ge 1-\frac{500}{6k^2}$$

$$\Rightarrow P[100-k\le S\le 100+k]\ge 1-\frac{500}{6k^2}$$

Choosing k = 20 we have

$$P[80\le S\le 120]\ge 1-\frac{500}{6x20^2}$$

$$\Rightarrow P[80\le S\le 120]\ge\frac{19}{24}$$

(3) Use Chebyshev's inequality to determine how many times a fair coin must be

tossed in order that the probability will be at least 0.90 that the ratio of the observed number of heads to the number of tosses will lie between 0.4 and 0.6 .

Sol. Let X denotes the number of heads obtained when a fair coin is tossed n times. Then proportion of heads in n tosses is X/n .Also X ~B(n,p) with p = ½ (since the coin is unbiased)

Then $E\left(\frac{x}{n}\right) = p = 1/2$ and $\text{Var}\left(\frac{x}{n}\right) = \frac{pq}{n} = \frac{1}{4n}$

So using Chebyshev's inequality for proportion of heads i.e. $\frac{x}{n}$, we get

$$P\left[\left|\frac{x}{n} - p\right| \geq k\right] \leq \frac{1}{4nk^2}$$

$$P\left[\left|\frac{x}{n} - p\right| < k\right] \leq 1 - \frac{1}{4nk^2}$$

Since we want the proportion of heads i.e. $\frac{x}{n}$ to lie between 0.4 and 0.6, so we have

$$P\left[\left|\frac{x}{n} - 0.5\right| < k\right] \geq 1 - \frac{1}{4nk^2}$$

$$\Rightarrow P\left[0.5 < k < \frac{x}{n} < 0.5 + k\right] \geq 1 - \frac{1}{4nk^2}$$

Choosing k =0.1 we have

$$\Rightarrow P\left[0.5 < \frac{x}{n} < 0.6\right] \geq 1 - \frac{1}{4n(0.1)^2}$$

Since we want this probability to be 0.9, so

$$1 - \frac{1}{4n(0.1)^2} = 0.9$$

$$\Rightarrow 0.10 - \frac{1}{0.04n} \Rightarrow n = 250$$

4. Does there exist a variate X for which

$$P[\mu_x - 2\sigma \le X \le \mu_x + 2\sigma\,]{=}0.6$$

Sol. We have, by Chebyshev's inequality

$$P[|X - \mu| < c] \ge 1 - \frac{\sigma^2}{c^2}$$

Consider $P[\mu_x - 2\sigma \le X \le \mu_x + 2\sigma] = P[|X - \mu_x| \le 2\sigma]$

Compairing it with Chebyshev's inequality, we get c = 2σ

So $P[\mu_x - 2\sigma \le X \le \mu_x + 2\sigma] = P[|X - \mu_x| \le 2\sigma] \ge 1 - \frac{\sigma^2}{4\sigma^2}$

Hence $P[\mu_x - 2\sigma \le X \le \mu_x + 2\sigma] \ge 1 - \frac{1}{4}$ i.e.

$$P[\mu_x - 2\sigma \le X \le \mu_x + 2\sigma] \ge \frac{3}{4}$$

Thus the lower bound for the probability is 0.75 ,so there does not exist a variate X for which

$$P[\mu_x - 2\sigma \le X \le \mu_x + 2\sigma] = 0.6$$

5. If f(X) = X^2 and $E(x)^2$ exists then $E(X)^2 \ge E(X)^2$

This is because if f(X) = X^2 then f`(x)=2 $\ge$ 0 hence f(X) is convex so by Jensen's inequality $E(X^2) \ge E(X)^2$

6. If f(X) = 1/X, X > 0 then

$f^{11}(x) = \frac{2}{x^3} > 0$ for $x > 0$, thus f(x) is convex

Hence $E = \left(\frac{1}{x}\right) \ge \frac{1}{E(X)}$

7. If f(X) =log X , X > 0 then

$f^{11}(x) = \frac{-1}{x^2} < 0$ so f(x) is concave

Hence $E[\log X] \leq [E(X)]$

11.9 SUMMARY

In this section we have dealt with some important moments inequalities viz. Chebyshev's inequality, Markov's inequality and Jensen's inequality. The significance of the first two inequalities lies in the fact that the actual distribution of the underlying random variable is not needed only the existence of lower order moments is required whereas Jensen's inequality is one which deals with convex functions.

11.10 SELFASSESSMENT:

(1) For a geometric distribution $p(x) = 2^{-x}$; x = 1, 2, 3 .... Show that Chebyshev's inequality gives

$$P[| X - 2 | \leq] \geq \frac{1}{2}$$

2. If X is a random variable such that $E(X) = 3$ and $E(X^2) = 13$, use Chebyshev's inequality to determine a lower bound for $P(-2 < X < 8)$.

(3) Two unbiased dice are thrown simultaneously .If denotes X the sum of the number showing up, prove that

$$P[| X - 7 | \leq 3] \geq \frac{35}{34}$$

11.11 FURTHER READING:

1. Modern Probability Theory,B.R.Bhatt
2. An Introduction to Probability Theory and its Application,w.Feller
3. Introduction to the Theory of Statistics: A M Mood ,F A Graybill and D C Boes
4. Fundamentals of Mathematical Statistics : S C Gupta and V K Kapoor

****

UNIT - IV **LESSON -12**

REGRESSION

STRUCTURE

12.1 INTRODUCTION:

After establishing the fact of correlation between two variables, it will be one's natural curiosity to know the extent to which one variable varies in response to a given variation in the other variable, i.e. one is interested to know the nature of relationship between two variables. This is done with the help of regression.

The term 'regression' was used by Galton in connection with the inheritance of stature. Galton found that the sons of fathers who deviate x inches from the mean height of all fathers themselves deviate from the mean height of all sons by less than x inches i.e. what Galton called a "Regression to Mediocrity".

In general the idea ordinarily attached to the word 'regression' does not touch upon this connotation and it should be regarded merely as a convenient term.

Regression is the estimation or prediction of unknown values of one variable from unknown values of another variable. For example, if we know that the height and weight are correlated we may be interested in finding out the most probable heights at certain given weights or the most probable weight at different heights. This can be done with the help of regression. The line which describes the average relationship between two variables is termed as lines of regression. The equations describing the regression lines are called regression equations.

12.2 OBJECTIVES:

On going through this unit reader will know how to

- Apply Simple Linear Regression Models
- Fit a Regression Line to the given data
- Obtain an equation that you can use for prediction and estimation purpose.

12.3 LINEAR AND NON- LINEAR REGRESSION:

If the given bivariate data are plotted on the graph paper, the points so obtained on the scatter diagram will more or less concentrate round the curve, called the 'curve of regression'. If the regression curve is a straight line, we say that there is linear regression between the variables under study. The equation of such a curve is the equation of a straight line i.e. a first degree equation in the variables X and Y. The Linear Regression equation Y upon X is Y = a + bX, and the linear Regression equation of X upon Y is X= a + b Y.

From the equation Y upon X we can estimate the most probable value of Y for given value of X. Similarly from the equation X upon Y we can estimate the most probable value of X for given value of Y.

However, if the curves of regression are not straight line, the regression is termed as

Curved or Non-Linear Regression.

12.4 DEFINITION:

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

12.5 WHY DO WE HAVE TWO REGRESSION LINES?

For two variables we have two regression equations–one gives the best estimate of one variable when the other is exactly known or given. For example if husband are denoted by X and wives by Y, the two equations will be

1. $X = a + bY$ and
2. $Y = a + bX$

From the first equation we will be in a position to obtain the most probable age of husbands for a given age of wife. The second equation gives the most probable estimate of the age of wife when the husband is given.

Corresponding to the two equations of regression two lines can be drawn of the graph paper and the extent of relationship can be found out from them. They are called regression line of lines of best-fit. They cut each other at the point of average of X and Y. The lower the value of the coefficients of correlation the farther would be the regression lines from each other. When the value of r is ±1 there will be only one regression line. The single straight line so obtained will give the required values of both the variables.

If we are given X and Y values we can fit a regression equation of the type $Y = a + bX$, Regression equation of Y on X. Here Y is dependent variable, a and b are two unknown constants and X is independent variables. In order to find out the values of a and b we have to solve two normal equations i.e.

$EY = Na + bEX$

$E(XY) = aE(X) + bE(X^2)$

We can also fit a regression equation of X on Y which will have general form–

$X = a + bY$

Here X is dependent variable, a and b are two unknown constants and Y is independent variables.

The values of a and b can be obtained as follows:

X = Na + b ∑ Y

XY = a∑Y + b∑Y^2

ANOTHER METHOD

Regression equations of X on Y will be written as follows

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

Where X and Y are variables under study, X is dependent and Y is independent variables. $\overline{X}$ and $\overline{Y}$ refer to the arithmetic mean of X and Y series.

r = coefficient of correlation.

σ_x and σ_y refer to the standard deviation of X and Y series.

Regression equations of Y on X will take the following form:

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

X is independent and Y is dependent variables. The quantity $r\frac{\sigma_x}{\sigma_y}$ in the first equation above is known as the regression coefficient of X on Y and is denoted by the symbol b_{xy}. The quantity $r\frac{\sigma_y}{\sigma_x}$ in the second equation above is known as the regression coefficient of Y on X and is denoted by the symbol b_{yx}.

12.6 SIMPLE REGRESSION LINE AND REGRESSION COEFFICIENT:

In this section, we describe a method of fitting a straight line equation to the given bivariate data.

The line of regression of Y on X is the line which gives the best estimate for the value of Y for any specified value of X.

Similarly, the line of regression of X on Y is the line which gives the best estimate for the value of X for any specified value of Y.

The term best fit is interpreted in accordance with the principle of least squares which consist in minimizing the sum of square of the residuals or error of the estimates i.e. the deviation between the given observed value of the variable and their corresponding estimated value as given by the line of best fit.

Minimizing the sum of the squares of the errors parallel to Y -axis gives, the equation of the line of regression of Y on X and minimizing the sum of the squares of the errors parallel to X-axis gives the equation of the line of regression of X on Y.

Let us consider that in a bivariate distribution $\left(X_i, Y_i\right), i = 1,2,3,........n$. Y is dependent and X is independent variable. Let the line of regression of Y on X be

$$Y= a + bX$$

Now according to the principle of least squares the normal equation for estimating a and b are:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \qquad (1)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_1^2 \qquad (2)$$

Dividing equation (1) by n, we get

$$\frac{1}{n}\sum_{i=1}^{n} y_i = a + b\frac{\sum_{i=1}^{n} x_i}{n}$$

$$\Rightarrow \qquad \overline{y} = a + b\overline{x} \qquad (3)$$

Thus, the line of regression passes through $\overline{x}$ and $\overline{y}$.

Also divide equation (2) by n, we get

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = a\frac{\sum_{i=1}^{n} x_i}{n} + b\frac{\sum_{i=1}^{n} x_1^2}{n} \qquad (4)$$

We know that

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$$

$$\Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma_x^2 + \overline{x}^2 \qquad (5)$$

Also $\quad r(x, y) = \dfrac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}$

$$r(x, y) = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}}{\sigma_x \sigma_y}$$

$$\Rightarrow \quad r\sigma_x \sigma_y = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$$

$$\Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} x_i y_i = r\sigma_x \sigma_y + \overline{x}\,\overline{y} \qquad (6)$$

Now substituting the equation (5) and (6) in equation (4), we have

$$r\sigma_x \sigma_y + \overline{x}\,\overline{y} = a\overline{x} + b\left[\,\sigma_x^2 - \overline{x}^2\,\right] \qquad (7)$$

Multiply equation (3) by $\overline{x}$ and subtracting from (7) we get

$$\Rightarrow \quad b = \frac{r\sigma_x \sigma_y}{\sigma_x^2} = \frac{r\sigma_y}{\sigma_x}$$

Since b is the slope of the line of regression of Y on X and since line of regression passes through $(\overline{x}, \overline{y})$ and its equation is

$y - \overline{y} = \dfrac{r\sigma_y}{\sigma_x}(x - \overline{x})$ is the line of regression of Y on x .

Similarly the equation of line of regression of X on Y is

$$x - \bar{x} = \frac{r\sigma_y}{\sigma_x}(y - \bar{y})$$

12.7 USEFULNESS:

The study of regression is very useful in many types of analysis. By its study we are able to obtain most probable values of one series for given values of the other related series. If we know that two series relating to price and supply are correlated we can find out what would be the effect on price if the supply of commodity was increased or decreased to a particular level.

The utility of regression is very great in physical services where the data are generally in functional relationship. There it is always possible to exactly calculate the value of one variable for a given value of the other variable by studying their regression. Thus the primary use of the regression equation is to describe the nature of the relationship between the two variables and to show the rates of change in one factor in terms of another.

<u>EXAMPLE</u>: Given

$$\sum X = 56, \quad \sum Y = 40, \quad \sum X^2 = 524,$$

$$\sum Y^2 = 256 \quad \sum XY = 364 \quad N = 8$$

Find the regression equation of X on Y.

SOLUTION: Regression equation of X on Y:

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

$$\bar{X} = \frac{\sum X}{N} = \frac{56}{8} = 7; \quad \bar{Y} = \frac{\sum Y}{N} = \frac{40}{8} = 5$$

$$r\frac{\sigma_x}{\sigma_y} \quad \text{or} \quad b_{xy} = \frac{\sum XY - N\bar{X}\bar{Y}}{\sum X^2 - N\bar{Y}^2}$$

$$= \frac{364 - (8 \times 7 \times 5)}{256 - 8(5)^2} = 1.5$$

X–7 = 1.5 (Y–5)

$\Rightarrow \quad X = -0.5 + 1.5Y.$

EXAMPLE: Calculate the regression of Y on X and predict the average value of Y when X is 9.

X:	3	6	5	4	4	6	7	5
Y:	3	2	3	5	3	6	6	4

SOLUTION:

Calculation of Regression Equation

	(X- $\bar{X}$)			(Y- $\bar{Y}$)		
	$\bar{X} = 5$			$\bar{Y} = 4$		
X	x	x^2	Y	y	y^2	xy
3	-2	4	3	-1	1	2
6	1	1	2	-2	4	-2
5	0	0	3	-1	1	0
4	-1	1	5	1	1	-1
4	-1	1	3	-1	1	1
6	1	1	6	2	4	2
7	2	4	6	2	4	4
5	0	0	4	0	0	0
SX=40	Sx=0	Sx^2=12	SY=32	Sy= 0	Sy^2=16	Sxy = 6

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} = \frac{6}{\sqrt{12 \times 16}} = 0.433$$

Regression equation of Y on X:

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}\left(X - \bar{X}\right)$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{32}{8} = 4; \qquad \bar{X} = \frac{\sum X}{N} = \frac{40}{8} = 5$$

$$r\frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum x^2} = \frac{6}{12} = 0.5$$

Substituting the values

$Y-4 = 0.5\ (X-5)$

$Y-4 = 0.5X - 2.5$ or $Y = 1.5 + 0.5X$

When X = 9 then

$Y = 1.5 + 0.5(9)$

$= 1.5 + 4.5$

$= 6.$

12.8 SELF ASSESSMENT:

1) What is regression analysis? Why are there two regression lines in case of bivariate series?

2) Given the following values of x and y:

x:	3	5	6	8	9	11
y:	2	3	4	6	5	8

find the equation of regression of (1) y on x (2) x on y.

3) From the following data, find out

a) Coefficient of correlation between the age of husbands & wives.

b) Two regression equations.

c) Most likely age of husbands when the wife age is 30.

d) Most likely age of wife when the husband age is 26.

Age of husband:	22	23	23	24	26	27	27	28	30	30
Age of wives:	18	20	21	20	21	22	23	24	25	26

4) If θ is the acute angle between the two regression lines, Prove that $Sin\theta \le 1 - r^2$.

****

UNIT – IV LESSON-13

REGRESSION COEFFICIENTS AND THEIR PROPERTIES

STRUCTURE

13.1 INTRODUCTION:

In the previous lesson, we have studied regression analysis, which is meant to determine the best functional relationship between a dependent variable with one or more related variable. In this lesson, regression coefficients and their properties have been discussed.

13.2 OBJECTIVES:

The main objectives of this lesson are

- To introduce the concept of Regression coefficients.
- To mention various properties of Regression coefficients.

13.3 REGRESSION COEFFICIENT:

'b' the slope of the line of regression of Y on X is also called the coefficient of regression of Y on X. It represents the increment in the value of dependent variable y corresponding to a unit change in the value of independent variable X.

More precisely, we write

b = Regression coefficient of Y on X

$$= \frac{\operatorname{cov}(X,Y)}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x}$$

Similarly, the coefficient of regression of X on Y indicates the change in the value of variable X corresponding to a unit change in the value of variable Y and is given by

b = Regression Coefficient of X on Y

$$= \frac{\operatorname{cov}(X,Y)}{\sigma_y^2} = r\frac{\sigma_x}{\sigma_y}$$

13.4 PROPERTIES OF REGRESSION COEFFICIENTS

1. Correlation Co-efficients is the geometric mean between the regression co-efficcients.

PROOF: The regression coefficients are

$r\frac{\sigma_y}{\sigma_x}$ is known as the regression coefficient of Y on X.

and $r\frac{\sigma_x}{\sigma_y}$ is known as the regression coefficient of X on Y.

Geometric mean between them

$$\text{G.M} = \sqrt{r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y}}$$

$$= \sqrt{r^2}$$

$$= r$$

= correlation coefficient.

2. If one of the Regression co-efficients is greater than unity numerically, the other

must be less than unity numerically.

PROOF: The two regression co-efficients of y on x

$b_{yx} = r\frac{\sigma_y}{\sigma_x}$ is known as the regression coefficient of y on x

and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ is known as the regression coefficient of x on y.

Let $b_{yx} > 1$

Then $\frac{1}{b_{yx}} < 1$ (1)

Since $b_{yx} . b_{xy} = r^2 \leq 1$ $\because -1 \leq r \leq 1$

$\because$ $b_{xy} \leq \frac{1}{b_{yx}} < 1$ 'using (1)

Similarly, if $b_{xy} > 1$, then $b_{yx} < 1$.

3. Arithmetic mean of the Regression Coefficient is greater than the Correlation coefficient r, provided $r > 0$.

PROOF: We have to prove that

$$\frac{1}{2}\left(b_{xy} + b_{yx}\right) \geq r$$

$$\frac{1}{2}\left[r\frac{\sigma_x}{\sigma_y} + r\frac{\sigma_y}{\sigma_x}\right] \geq r$$

$$\Rightarrow \frac{1}{2}r\left[\frac{\sigma_x^2 + \sigma_y^2}{\sigma_y \sigma_x}\right] \geq r$$

$$\Rightarrow \sigma_x^2 + \sigma_y^2 \geq 2\sigma_x\sigma_y$$

$\Rightarrow \sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y \geq 0$

$\Rightarrow \left(\sigma_x^2 - \sigma_y^2\right) \geq 0$

4. Regression Coefficients are independent of origin but not of scale.

PROOF: Let U = $\dfrac{X-a}{h}$ and V = $\dfrac{Y-b}{k}$

$\Rightarrow$ x = a + Uhand Y = b + Vk

Where a, b, h and k are constants.

We have to prove that $b_{xy} \neq b_{uv}$

Since X = a + hu (1)

$\Rightarrow$ E(X) = a + h E (U) (2)

and Y = b + v k (3)

$\Rightarrow$ E(Y) = b + k E (v) (4)

Subtract equation (2) from (1) and (4) from (3), we get

X – E(X) = h [U - E (U)] (5)

and Y – E(Y) = k [V - E (V)] (6)

$\therefore$ $\operatorname{cov}(X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$

= E [h [U – E (U)]. k [V – E (V)]]

= hk E [[U – E (U)] [V – E (V)]]

= hk cov (U, V) (7)

and

σ_x^2 = E[X–E(X)]^2

= E [h {U–E (U)}]^2

= h^2 σ_u^2 (8)

σ_y^2 = E [Y–E(Y)]^2

= E [k {V–E (V)}]^2

= k^2 σ_v^2 (9)

Therefore, $b_{xy} = \dfrac{\text{cov}(X,Y)}{\sigma_y^2}$

$$= \frac{\text{hk}\,\text{cov}(U,V)}{k^2\sigma_v^2}$$

$$= \frac{\text{h}\,\text{cov}(U,V)}{k\sigma_v^2}$$

$$= \frac{h}{k} b_{uv} \qquad (10)$$

Similarly,

$$b_{yx} = \frac{\text{cov}(X,Y)}{\sigma_x^2}$$

$$= \frac{\text{hk}\,\text{cov}(U,V)}{h^2\sigma_u^2}$$

$$= \frac{\text{k}\,\text{cov}(U,V)}{h\sigma_u^2}$$

$$= \frac{k}{h} b_{vu} \qquad (11)$$

Thus, from (10) and (11), we get

$b_{xy} \neq b_{uv}$ and $b_{yx} \neq b_{vu}$.

5. The correlation coefficient and the two regression co-efficients have same sign.

PROOF: Regression coefficient of y on x = $b_{yx} = r\dfrac{\sigma_y}{\sigma_x}$

Regression coefficient of x on y = $b_{xy=}\ r\dfrac{\sigma_x}{\sigma_y}$

Since σ_x and σ_y are both positive.

Therefore, b_{yx}, b_{xy} and r have same sign.

6. Angle between two lines of Regression.

PROOF: Equations to the lines of regression of y on x

$$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \text{ and}$$

Equations to the lines of regression of x on y

$$x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

Their slopes are $m_1=r\frac{\sigma_y}{\sigma_x}$ and $m_2=r\frac{\sigma_x}{\sigma_y}$ respectively

$$\therefore \tan\theta = \left|\frac{m_2-m_1}{1+m_2m_1}\right|$$

$$= \left|\frac{r\frac{\sigma_x}{\sigma_y}-r\frac{\sigma_y}{\sigma_x}}{1+\frac{\sigma_y^2}{\sigma_x^2}}\right|$$

$$= \left|\frac{1-r^2}{r}.\frac{\sigma_y}{\sigma_x}\frac{\sigma_x^2}{\sigma_x^2+\sigma_y^2}\right|$$

$$= \frac{1-r^2}{|r|}.\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

Since $r^2\leq 1$ and σ_x and σ_y are positive.

Hence $\tan\theta = \frac{1-r^2}{|r|}.\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$

Case-1 when r = 0, $\theta = \frac{\pi}{2}$

Therefore, the two lines of regression are perpendicular to each other.

Case-2 when r = ± 1, $\theta = 0$ or π

Hence the lines of regression coincide and there is perfect correlation between the two variates x and y.

EXAMPLE: If the two regression co-efficients are 0.8 and 0.2, what would be the value of co-efficient of correlation.

SOLUTION: Since the co-efficient of of correlation is the G.M between the regression co-efficients.

$\therefore \quad r^2 = 0.8 \times 0.2 \quad = 0.16$

Since $r = \pm 0.4$

Since r has same sign as the two regression co-efficients

$\therefore \quad r$ must be +ve.

Hence r = 0.4.

EXAMPLE: Given

$n = 5, \quad \overline{X} = 10, \quad \overline{Y} = 20, \quad \sum(X-4)^2 = 100, \quad \sum(Y-10)^2 = 160,$
$\sum(X-4)(Y-10) = 80.$

Find the two regression coefficients and hence, the coefficient of correlation.

SOLUTION: Let U = X–4 and V = Y–10.

Then we are given

$n = 5, \ \overline{X} = 10, \ \overline{Y} = 20, \ \sum(X-4)^2 = \sum U^2 = 100, \ \sum(Y-10)^2 = \sum V^2 = 160,$
$\sum(X-4)(Y-10) = \sum UV = 80$

Also U = X–4

$\Rightarrow \overline{U} = \overline{X} - 4$ = 10–4 = 6

$\Rightarrow \sum U = n\overline{U} = 5 \times 6$ = 30.

And V = Y – 10

$$\Rightarrow \; \overline{V} = \overline{Y} - 10 = 20 - 10 = 10$$

$$\Rightarrow \; \sum V = n\overline{V} = 5 \times 10 = 50$$

Since, the regression coefficients are independent of the change of origin, the regression coefficients are given by

$$b_{YX} = b_{VU} = \frac{n\sum UV - (\sum U)(\sum V)}{n\sum U^2 - (\sum U)^2}$$

$$= \frac{5(80) - (30)(50)}{5(100) - (30)^2}$$

$$= \frac{11}{4}$$

$$\therefore \; b_{XY} = b_{UV} = \frac{n\sum UV - (\sum U)(\sum V)}{n\sum V^2 - (\sum V)^2}$$

$$= \frac{5(80) - (30)(50)}{5(160) - (50)^2}$$

$$= \frac{11}{17}.$$

The correlation coefficient ‘r’ is given by:

$$r_{XY} = \pm\sqrt{b_{YX}.b_{XY}}$$

$$= \pm\sqrt{\frac{11}{4} \times \frac{11}{17}}$$

$$= \pm 1.33 \cdot$$

But, since the regression coefficients are positive, we take:

$$r_{XY} = 1.33 > 1$$

Which is impossible, since $|r| \le 1$.

13.5 SELF ASSESSMENT:

1) Is the following statement true? Give reasons.

   $40x - 18y = 5$ and $8x - 10y + 6 = 0$

   are respectively the regression equations of y on x and x on y.

2) Find the regression coefficients b_{xy} and b_{yx} from the following data:

   $\sum X = 30$, $\sum Y = 42$, $\sum X^2 = 184$,

   $\sum Y^2 = 318$ $\sum XY = 199$ $n = 6$

3) Obtain the regression lines of y on x and x on y from the following table and estimate the blood pressure when age is 45 years.

Age in Years:	57	43	73	37	64	48	56	50	39	43
Blood Pressure (y):	148	127	161	119	150	129	151	146	116	141

****

UNIT - IV LESSON -14

LEAST SQUARE METHOD

STRUCTURE

14.1 INTRODUCTION:

The principle of least squares is the most popular and widely used method of fitting mathematical functions to a given set of data and best ways of obtaining trend values. With this method, a straight line is fitted. This line is called the line of the best fit. It is a line from which the sum of the deviations of various points on either side is equal to zero.

This method gives us a set of equations containing unknown constants i.e. parameters are called of the Normal equations and serve to determine these unknown constants. This special criterion of the 'best fit' of the data is called the 'principle of least square.' The method of finding the unknown parameters by the principle of least squares is called the method of least squares.

14.2 OBJECTIVES:

The main objectives of this lesson to know:

- about the linear equations.
- about the normal equations.
- the values of constants by using normal equations.
- about the future values how to be projected.

14.3 THE METHOD OF LEAST SQUARES:

The method of least squares is the most popular method of fitting a trend line to the data. The method of least squares is a mathematical device which places a line through a series of plotted points in such a way that the sum of squares of the deviation of the actual pints above and below the trend line is at the minimum. The method of least squares gives us what is known as the line of best fit. It is a line from which the sum of deviations of various points on either side is equal to zero.

In other words, if we sum up the positive and negative deviations on either side of the line of best fit the sum will be zero. This being so the sum of the squares of these deviations obtained will be the least as compared to the sum of squares of the deviations obtained by using other lines. It is on account of this fact that this method is known as the method of least squares.

<u>MOST PLAUSIBLE SOLUTION OF A SYSTEM OF LINEAR EQUATION:</u>

Suppose we have a number of independent linear equations in n unknowns $x_1, x_2, ...x_n$,

$$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2$$

$$.................................$$

$$a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = b_n$$

or

$$\sum_{j=1}^{n} a_{ij}x_i = b_i , \qquad i = 1,2,........., m$$

Where $a's$ and $b's$ are constants. If $m = n$, we can, in general find a unique set of values satisfying the given system of equations. However, if $m > n$ that is if the number of equations is greater than the number of unknowns, no such solutions exist. We, therefore, try to find those values of $x_1, x_2, ...x_n$ which will satisfy the given system of equations as nearly as possible.

The principal of least squares asserts that these values are those which make S a minimum where

$$S = \sum_{i=1}^{m} (a_{i1}x_1 + a_{i2}x_2 + ... + a_{in}x_n - b_i)^2$$

$$= \sum_{i=1}^{m} E_i^2$$

where

$$E_i = a_{i1}x_1 + a_{i2}x_2 + ... + a_{in}x_n - b_i$$

We shall call these values the best or most plausible values in the least square sense.

By a well-known theorem of differential calculus the extreme values of the function

$$f = f(x_1, x_2, ...x_n)$$

are given by $\dfrac{\partial f}{\partial x_1} = \dfrac{\partial f}{\partial x_2} = ... = \dfrac{\partial f}{\partial x_n} = 0$

provided that the partial derivatives exist.

Applying this theorem, S will have a maximum or minimum for those values of the unknowns $x_j\,(j=1,2,........., n)$ which satisfy the following n equations:

$$\frac{\partial S}{\partial x_j} = 0, \quad j = 1,2,..........., n$$

i.e. $$\sum_{i=1}^{m} a_{i1}E_i = 0, \ \sum_{i=1}^{m} a_{i2}E_i = 0, \ ...., \sum_{i=1}^{m} a_{in}E_i = 0 \qquad (1)$$

These n equations are called the normal equations corresponding to the equations (1). They can be solved like simultaneous equations for the n unknowns $x_1, x_2, \ldots\ldots\ldots\ldots x_n$. If the common rank of the matrix of coefficients and the augmented matrix is n, we can get a unique solution for $x_1, x_2, \ldots\ldots\ldots\ldots x_n$.

EXAMPLE: Form normal equations and hence find the most plausible values of x and y from the following equations:

$$x + y = 3.01,\ 2x - y = 0.03,\ x + 3y = 7.03,\ 3x + y = 4.97.$$

SOLUTION:

Here $\quad S=(x + y - 3.01)^2 + (2x - y - 0.03)^2 + (x + 3y - 7.03)^2 + (3x + y - 4.97)^2$

The normal equations are

$$\frac{\partial S}{\partial x}=0 \text{ and } \frac{\partial S}{\partial y}=0$$

$$(x + y - 3.01) + 2(2x - y - 0.03) + (x + 3y - 7.03) + 3(3x + y - 4.97) = 0$$

and

$$(x + y - 3.01) - (2x - y - 0.03) + 3(x + 3y - 7.03) + (3x + y - 4.97) = 0$$

i.e., $\quad 15x + 5y = 25$

and $\quad 5x + 12y = 29.01$

Solving these equations, we get $x = 0.9997$, $y = 2.0009$ approx.

14.4 MERITS OF THE LEAST SQUARES:

1. This method is not affected by the personal prejudice and bias of the computer since it is based on a mathematical equation.
2. It is possible to obtain trend values for all the items of the series.
3. This method gives us the most satisfactory results because the sum of the positive and negative deviations of the actual plotted points from the trend line is zero and the sum of squares of these deviations is the least.

14.5 DEMERITS OF THE LEAST SQUARES

1) The method is quite tedious and time consuming as compared with other methods.

2) Even if a single item is added to the series, all calculations to be done afresh.

3) The most serious limitation of this method is the determination of the type of the trend curve to be fitted.

14.6 PRINCIPAL OF LEAST SQUARES:

The principal of least squares provides us an analytical or mathematical device to obtain an objective to fit the trend of the given time series. It is the most popular and widely used method of fitting mathematical functions to a given set of observations. The curve fitted by least squares method is the only technique which enables us to obtain the rate of growth per annum, for yearly data, if linear trend is fitted. The various types of lines that may be used to describe the given data in practice are:

1) A Straight line : $y = a + bx$

2) Second degree parabola: $y = a + bx + cx^2$

14.7 FITTING OF LINEAR TREND

Let the straight line trend between the given time series values (y) and time (t) be gien by the equation:

$$y = a + bt \qquad (1)$$

Then for any given time 't', the estimated value y_e of y as given by this equation is :

$$y_e = a + bt \qquad (2)$$

by using the principal of least squares consist in estimating the values of a and b in (1) so that the sum of the squares of errors of estimate

$$E = \Sigma(y-y_e)^2$$
$$= \Sigma\ (y-a-bt)^2 \qquad (3)$$

is minimum, the summation being taken over given values. This will be so if:

$$\frac{\partial E}{\partial a} = 0 = -2\Sigma\ (y-a-bt)$$

and (4)

$$\left.\frac{\partial E}{\partial b} = 0 = -2\Sigma t(y\text{-}a\text{-}bt)\right\}$$

which, gives the normal equations or least square equations for estimating a and b as

$$\left.\begin{array}{l}\Sigma y = na + b\Sigma t \\ \Sigma ty = a\Sigma t + b\Sigma t^2\end{array}\right\} \quad (5)$$

where n is the number of observations.

Solving equations (4) and (5) for a and b and substituting these values in equation (1), we finally get the equation of straight line trend.

14.8 FITTING OF SECOND DEGREE PARABOLA BY THE METHOD OF LEAST SQUARES

Let the second degree parabolic trend be given by the equation:

$$y = a + bt + ct^2 \quad (1)$$

Then for any given value of t, the trend value is given by

$$y_e = a + bt + ct^2$$

Thus, if y_e is the trend value corresponding to an observe value y, then according to the prioncipal of least squares we have to obtain the value of a, b and c in equation (1) so that

$$E = \Sigma\ (y\text{-}y_e)^2$$

$$= \Sigma\ (y\text{-}a\text{-}bt\text{-}\ ct^2)^2$$

is minimum for variations in a, b and c. Thus, the normal or least squares equations for estimating a, b and c are given by:

$$\frac{\partial E}{\partial a} = 0 = -2\Sigma\ (y\text{-}a\text{-}bt\text{-}\ ct^2)$$

and (2)

$$\left.\begin{aligned} \frac{\partial \mathrm{E}}{\partial \mathrm{b}} &= 0 = -2\Sigma t(y-a-bt-\ ct^2) \\ \frac{\partial E}{\partial c} &= 0 = -2\Sigma t^2(y-a-bt-\ ct^2) \end{aligned}\right\}$$

$$\Rightarrow \quad \left.\begin{aligned} \Sigma y &= na + b\Sigma t + c\Sigma t^2 \\ \Sigma ty &= a\Sigma t + b\Sigma t^2 + c\ \Sigma t^3 \\ \Sigma t^2 y &= a\Sigma t^2 + b\Sigma t^{3\,+}\ c\Sigma t^4 \end{aligned}\right\} \qquad (3)$$

Solving equations (3) for a, b and c and substituting these values in equation (1), we finally get the Parabolic curve.

<u>EXAMPLE</u>: Fit a straight line to the following data:

t:	0	1	2	3	4
y:	1	1.8	3.3	4.5	6.3

<u>SOLUTION</u>: Let the straight line to be fitted to the data be

$$y = a + bt$$

Then the normal equations are

$$\Sigma y = na + b\Sigma t$$

$$\Sigma ty = a\Sigma t + b\Sigma t^2$$

t	y	ty	t^2
0	1	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16
Total 10	16.9	47.1	30

Here $n = 5$, $\Sigma t = 10$

$\Sigma y = 16.9$, $\Sigma ty = 47.1$, $\Sigma t^2 = 30$

On substituting these values, the normal equations become

$$16.9 = 5a + 10b \qquad (1)$$

$$47.1 = 10a + 30b \qquad (2)$$

Solving (1) and (2), we get

$$a = 0.72$$

$$b = 1.33$$

Hence the equation of the line of best fit is

$$Y = 0.72 + 1.33t.$$

<u>EXAMPLE</u>: Fit a second degree parabola to the following data:

t:	0	1	2	3	4
y:	1	5	10	22	38

SOLUTION: Let the curve of best fit be

$$y = a + bt + ct^2$$

Then the normal equations are

$$\Sigma y = na + b\Sigma t + c\Sigma t^2$$

$$\Sigma ty = a\Sigma t + b\Sigma t^2 + c\ \Sigma t^3$$

$$\Sigma t^2 y = a\Sigma t^2 + b\Sigma t^{3\,+}\ c\Sigma t^4$$

t	y	ty	t^2	$t^2\ y$	t^3	t^4
0 1 2 3 4	1 5 10 22 38	0 5 20 66 152	0 1 4 9 16	0 5 40 198 608	0 1 8 27 64	0 1 16 81 256
10	76	243	30	851	100	354

Here n = 5

The normal equations become

$$76 = 5a + 10b + 30c \qquad (1)$$

$$243 = 10a + 30b + 100c \qquad (2)$$

$$851 = 30a + 100\ b + 354\ c \qquad (3)$$

after solving the equations, we get

$a = 1.42$, $b = 0.26$, $c = 2.21$.

Hence the equation of the curve is

$$y = 1.42 + 0.26t + 2.21t^2.$$

14.9 SELF ASSESSMENTS:

1) Explain the Principle of least squares. Also discuss their merits and demerits?

2) Find the most plausible values of x and y from the following equations.

$$x - 5y + 4 = 0 \qquad 2x - 3y + 5 = 0$$
$$x + 2y - 3 = 0 \qquad 4x + 3y + 1 = 0$$

3) Fit an equation of the form $Y = a + bX + cX^2$ to the data given below:

X:	1	2	3	4	5
Y:	25	28	33	39	46

4) Fit a straight line to the following data:

X:	1	2	3	4	6	8
Y:	2.4	3	3.6	4	5	6

****

UNIT - IV LESSON - 15

LOGARITHMIC AND EXPONENTIAL CURVES

STRUCTURE:

15.1 INTRODUCTION:

The method of line fitting by the Principle of Least squares is used quite often in the trend analysis, particularly when one is interested in making projections for the future times.

Sometimes it may happen that the original data is not in a linear form but can be reduced to linear form by some simple transformation of variables. In this lesson we study the fitting of curves reducible to polynomials by transformations.

15.2 OBJECTIVES:

On reading this lesson you should be able to know:

- About non- linear curves i.e. power curve.
- Exponential curves

15.3 FITTING OF CURVES:

1) **Fitting of a power curve:** $Y = a X^b$

Suppose we want to fit a power curve $Y = a X^b$ .....(1) to a set of n pairs of observations (X_i, Y_i), i = 1 ,2,...........n

Taking logarithmic of equation (1) both sides we have

Log Y = Log a + b Log X

$\Rightarrow$ U = A + bV, Where U=Log Y, A=Log a and V=Log X

This is a linear equation in U and V .Thus, the normal equations for estimating A and b are

$$\sum U = nA + b\sum V$$

$$\sum UV = A\sum V + b\sum V^2$$

On simplification of these equations, we get the values of A and b and consequently, we get

a = antilog (A)

Substituting the values of a and b in equation (1), we get the fitted power curve.

2) **Fitting of exponential curve of the form**: $Y= ab^X$

Suppose we want to fit an exponential curve of the type

$$Y = ab^X \qquad (1)$$

to a set of n pairs of observations (X_i,Y_i), i= 1,2,3............n

Taking logarithmic both sides of equation (1), we get

Log Y = Log (ab^X)

= Log a+ X Log b

$$U = A+ BX, \qquad (2)$$

Which is linear equation in X and U

Where U = Log Y, A= Log a, B= Log b

The normal equations for estimating A and B are

$$\sum U = nA + B\sum X$$

$$\sum UX = A\sum X + B\sum X^2$$

On simplification of these equations, we get the values of A and B and consequently a= antilog (A), b = antilog (B)

Finally, substituting the values of a and b in equation (1), we get the fitted an exponential curve.

3) **Fitting of an exponential curve of the form**: $Y= a_e^{bX}$

Suppose we want to fit an exponential curve of the form

$$Y= a_e^{bX} \qquad (1)$$

to a set of n pairs of observations (X_i,Y_i) ; i =1,2,……..n

Taking logarithm on both sides of (1), we get

$$\log Y = \log (a_e^{bX})$$

$$\log Y = \log a + (b \log e) X$$

$$U= A +B X$$

Where U=log Y, A= log a and B= b loge

This in linear equation in X and U

The normal equations for estimating A and B are

$$\sum U = nA + B\sum X$$

$$\sum UX = A\sum X + B\sum X^2$$

On simplification of these equations, we get the values of A and B and consequently

$a = \text{antilog}\,(A),\ b = \dfrac{B}{\log e}$

Finally, substituting the values of a and b in equation (1), we get the fitted an exponential curve.

EXAMPLE: Fit an exponential curve of the form $Y = ab^X$ to the following data:

X:	1	2	3	4	5	6	7	8
Y:	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

SOLUTION: The exponential curve is

$$Y = ab^X$$

Normal equations are

$$\sum U = nA + B\sum X$$

$$\sum UX = A\sum X + B\sum X^2$$

Where $U = \log Y$, $A = \log a$ and $B = \log b$

X	Y	U=log Y	UX	X^2
1	1.0	0.000	0.000	1
2	1.2	0.0792	0.1584	4
3	1.8	0.2553	0.7659	9
4	2.5	0.3979	1.5916	16
5	3.6	0.5563	2.7815	25
6	4.7	0.6721	4.0326	36
7	6.6	0.8195	5.7365	49
8	9.1	0.9590	7.6720	64
$\sum X$ =36	$\sum Y$=30.5	$\sum U$=3.7393	$\sum UX$ =22.7385	$\sum X^2$ =204

Substituting the values in the normal equations, we get

$$3.7393 = 8\text{ A} + 36\text{ B} \qquad (1)$$

$$22.7385 = 36\text{ A}+204\text{ B} \qquad (2)$$

Multiply equation (1) by 9 and (2) by 2 and then subtracting, we get

$$\Rightarrow \qquad -11.8233 = -\ 84\text{ B}$$

$$\Rightarrow \qquad \text{B} = \frac{11.8233}{84} = 0.1408$$

Put this value of 'B' in equation (1), we get

$$3.7393 = 8\text{ A} + 36\ (0.1408)$$

$\Rightarrow$ $8\,A = -\,1.3295$

$\Rightarrow$ $A = -0.1662 = \bar{1}.8338$

$\therefore$ b = antilog (B) and a = antilog (A)

$\Rightarrow$ b = 1.383 and a = 0.6821

Hence the required exponential curve is

$Y = 0.6821(1.383)^{X}$

EXAMPLE: Fit the exponential curve $Y = a_{e}^{bX}$ to the following data:

X:	0	2	4
Y:	5.012	10	31.62

SOLUTION: The equation to the curve can be written as:

$$\log Y = \log a + (b\ \log e)X$$

$$U = A + B\,X$$

The normal equations are

$$\sum U = nA + B\sum X$$

$$\sum UX = A\sum X + B\sum X^2$$

X	Y	U=logy	X^2	XU
0 2 4	5.012 10 31.62	0.7 1.0 1.5	0 4 16	0 2 6
$\sum X = 6$	$\sum Y = 46.632$	$\sum U = 3.2$	$\sum X^2 = 20$	$\sum UX = 8$

Substituting the values in the normal equations, we get

$$3.2 = 3(A) + 6B \qquad (1)$$

$$8 = 6(A) + 20B \qquad (2)$$

Multiply equation (1) by 2 then subtracting (2) from (1), we get

$$B = 0.2$$

Put the value of 'B' in equation (1), we get

$$3.2 = 3A + 6(0.2)$$

$$A = 0.666$$

$$a = \text{antilog}(A) = 4.642$$

$$b = \frac{B}{\log e}$$

$$= \frac{0.2}{\log(2.71828)}$$

$$= \frac{0.2}{0.4343} = 0.46$$

Hence the equation of the curve fitted is

$$Y = 4.642_e{}^{(0.46)X}$$

15.4 SUMMARY:

In this lesson, we read about

1) Exponential curves.

2) Power curve.

3) Normal equations.

15.5 SELF ASSESSMENT:

1) Fit an equation of the form $Y = aX^b$ to the following data

X:	4	5	6	7	8
Y:	146	174.8	209.4	250.8	300.6

2) In an experiment, in which the growth of duck weed under certain conditions was measured, the following results were obtained:

Weeds(X):	1	2	3	4	5	6	7	8	9
No. of friends (Y):	21	31	53	78	136	212	327	551	1053

Assuming that the relationship of the form $Y=a_e^{bX}$, find the best values of a and b by the method of least squares.

3) Fit an equation of the form $Y = aX^b$ to the following data:

X:	7	9	11	13	15
Y:	58	254	690	1462	2666

****

UNIT - IV **LESSON -16**

MULTIPLE CORRELATION COEFFICIENT

STRUCTURE:

16.1 INTRODUCTION:

So far we have considered correlation between two variates only. But often it is necessary to finds correlation between three or more variates. For example, the statures of men are influenced by those of all their ancestors, and the yield of grain is affected by the amounts of different fertilizers used.

Whenever we are interested in the combined influence of a group of variates upon a variate not included in the group, our study is that of multiple correlation.

In a multivariate population, the different variates may be mutually correlated and this correlation will, in general, be influenced by the other variates of the population. To study the relationship between any two variates, we have two methods.

Firstly, we may consider only those members of the observed data in which the other members have specified values.

Secondly, we may eliminate mathematically the effect of other variates on the two variates under study.

Thus the aim of the theory of multiple correlation is to know how far the dependent variable is influenced by the independent variables.

The correlation between two variates when the linear effect of the other variates in them has been eliminated from both is called partial correlation.

16.2 OBJECTIVES:

The main objective of this Lesson is

- to study the multiple regression and their variables
- to find the correlation of more than two variables.

16.3 MULTIPLE REGRESSION AND THEIR VARIABLES:

Let us consider a distribution involving three variables X_1, X_2 and X_3. Then the equation of plane of regression of X_1 on X_2 and X_3 is

$$X_1 = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3 \qquad (1)$$

Where X_1 is the dependent variable and X_2, X_3 are independent variables and $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ are constants whose values are to be estimated by the method of least squares for values of X_1, X_2 and X_3.

Since X_1 varies partially in the variation of X_2 and partially variation in the X_3. So we refer $b_{12.3}$ and $b_{13.2}$ as partial regression coefficient of X_1 on X_2 keeping X_3 as constant and of X_1 on X_3 keeping X_2 as constant respectively.

$b_{12.3}$ and $b_{13.2}$ are regression coefficients of first order.

The quantity $X_{1.23} = X_1 - b_{12.3} X_2 - b_{13.2} X_3$ is called error of estimate or residual.

The other two equations are b

When X_2 is dependent variable and X_1 and X_3 are independent variables is

$$X_2 = a_{2.13} + b_{21.3} X_1 + b_{23.1} X_3$$

Similarly, the regression of X_3 on X_1 and X_2 is

$$X_3 = a_{3.12} + b_{31.2} X_1 + b_{32.1} X_2$$

In the general case if n variables $X_1, X_2, X_3 \ldots\ldots\ldots\ldots X_n$, the equation of the plane of regression of X_1 on $X_2, X_3 \ldots\ldots\ldots\ldots X_n$ is

$$X_1 = b_{12.34.....n} X_2 + b_{13.24.......n} X_3 + \ldots\ldots\ldots\ldots\ldots\ldots + b_{1n.23........n-1} X_n$$

The error of the estimate is given by

$$X_{1.23.......n} = X_1 - b_{12.34.....n} X_2 - b_{13.24.......n} X_3 - \ldots\ldots\ldots\ldots - b_{1n.23........n-1} X_n .$$

16.4 PRIMARY AND SECONDARY SUBSCRIPTS:

We used a point separating the subscripts in case of some quantities like $x_{1.2}$, $\sigma_{2.1}$ etc. The subscripts preceding the point are called Primary Subscripts and the Subscripts following the point are called Secondary Subscripts.

The number of the Secondary Subscripts used to denote a specific quantity is referred to as Order. For instance

$x_{1.2}$ is a residual of first order.

r_{12} is the correlation coefficient of zero order.

16.5 MULTIPLE CORRELATION ANALYSIS:

In multiple correlation it is assumed that the dependent variable is related to a number of independent variables and the degree of association between the dependent variable and a number of independent variable taken together is measured .For instance, the yield of crop per acre say (X_1) depends upon quality of seeds (X_2), fertility of soil (X_3) fertilizer used (X_4), the irrigation facility (X_5), weather conditions (X_6) and so on

$$X_1 = X_2 + X_3 + X_4 + X_5 + X_6$$

Here, X_1 production is dependent variable and X_2, X_3, X_4, X_5, and X_6 factors are independent variables.

Whenever we are interested in studying the joints effects of a group of variables upon a variable not included in that group, our study is that of multiple correlations and multiple regressions. The multiple correlation coefficients is denoted by $R_{1.23.........n}$.

The subscripts of $R_{1.23.........n}$ shows that the relation studied is between the dependent variable X_1 and the set of independent variables X_2, X_3, X_4, X_5, and X_6.

In case of tri-variate the multiple correlation coefficients may be defined in terms of simple correlation coefficient as given below:

$\mathrm{R}_{1.23}=\sqrt{\dfrac{r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}}{1-r_{23}^2}}$, i.e. the multiple correlation between X_1 on the one hand and X_2 and X_3 on the other hand.

$\mathrm{R}_{2.13}=\sqrt{\dfrac{r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}}{1-r_{13}^2}}$, i.e. the multiple correlation between X_2 on the one hand and X_1 and X_3 on the other hand.

$\mathrm{R}_{3.12}=\sqrt{\dfrac{r_{13}^2+r_{23}^2-2r_{12}r_{13}r_{23}}{1-r_{12}^2}}$, i.e. the multiple correlation between X_3 on the one hand and X_1 and X_2 on the other hand.

The range of multiple correlation coefficient lies between 0 and 1 i.e. it is non-negative correlation.

EXAMPLE: Show that when $\mathrm{R}_{1.23}=1$ then $\mathrm{R}_{2.13}=1$

SOLUTION: we know that

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

Since R1.23 =1

$$\Rightarrow \quad 1=\frac{r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}}{1-r_{23}^2}$$

$$\Rightarrow \quad r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}=1-r_{23}^2$$

Also we know that

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

From equation (1), we have

$$r_{12}^2 + r_{13}^2 \;= 1 - r_{13}^2 + 2r_{12}r_{13}r_{23}$$

Substitute equation (3) in (2), we get

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

$$\Rightarrow \qquad R_{2.13}^2 = 1$$

EXAMPLE: Prove that $R_{1.23}^2 = b_{12.3}r_{12}\left(\frac{\sigma_2}{\sigma_1}\right) + b_{13.2}r_{13}\left(\frac{\sigma_3}{\sigma_1}\right)$

SOLUTION: $b_{12.3} = -\left(\frac{\sigma_1}{\sigma_2}\right)\left(\frac{w_{12}}{w_{11}}\right)$

$$\text{R.H.S} = -\left\{ \frac{\sigma_1}{\sigma_2} . \frac{w_{12}}{w_{11}} r_{12}\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1}{\sigma_3} . \frac{w_{13}}{w_{11}} r_{13}\left(\frac{\sigma_3}{\sigma_1}\right) \right\}$$

$$= -\left\{\frac{r_{12}\,(r_{23}r_{13} - r_{12}) + r_{13}\,(r_{21}r_{32} - r_{31})}{1 - r_{23}^2}\right\}$$

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

$$= R_{1.23}^2$$

=L.H.S

EXAMPLE: If $r_{ij} = corr(X_i, X_j)$, Show that

(1) $r_{12} + r_{23} + r_{31} \geq -\frac{3}{2}$

(2) $1 + 2r_{12}\,r_{23}\,r_{31} \geq r^2{}_{12} + r^2{}_{23} + r^2{}_{31}$

SOLUTION:

(1) Let $X^* = \dfrac{X - E(X)}{\sigma_x}$, then

$$E\left(X_1^* + X_2^* + X_3^*\right)^2 \geq 0$$

$$E\left(X_1^{*2} + X_2^{*2} + X_3^{*2} + 2X_1^* X_2^* + 2X_2^* X_3^* + 2X_1^* X_3^*\right) \geq 0$$

Since $E\left(X_i^*\right) = \text{var}\left(X_i^*\right) = 1$

$E\left(X_i^* X_j^*\right) = r_{ij}$, we get

$$1 + 1 + 1 + 2r_{12} + 2r_{23} + 2r_{31} \geq 0$$

$\therefore \qquad r_{12} + r_{23} + r_{31} \geq -\dfrac{3}{2}$

(2) Since $R_{1.23}^2 \leq 1$, we get

$$r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 1 - r_{23}^2$$

$\therefore \qquad 1 + 2r_{12}\, r_{23}\, r_{31} \geq r^2{}_{12} + r^2{}_{23} + r^2{}_{31}$

Proved.

<u>THEOREM</u>: Prove that multiple correlation coefficient is

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

<u>PROOF</u>:

In a trivariate distribution $X_{1,}$ X_2 and X_3 each has n observations. The multiple correlation coefficient of X_1 on X_2 and X_3 usually denoted by $R_{1.23}$, is the simple correlation coefficient between X_1 and the joint effect of X_2 and X_3 on X_1. The estimated value of the plane of regression of X_1 on X_2 and X_3 is

$$e_{1.23} = b_{12.3}\, X_2 + b_{13.2} X_3$$

We have $X_{1.23} = \mathrm{X}_1 - b_{12.3} X_2 - b_{13.2} X_3$

$\Rightarrow$ $X_{1.23} = \mathrm{X}_1 - e_{1.23}$

Or $e_{1.23} = \mathrm{X}_1 - X_{1.23}$

Since $\mathrm{X}_{1,}$ X_2 and X_3 are measured from their respective means, we have

$$E(X_1) = E(X_2) = E(X_3) = 0$$

$$\Rightarrow \quad E(X_{1.23}) = 0 \text{ and } E(e_{1.23}) = 0$$

$$\therefore \quad R_{1.23} = \frac{\operatorname{cov}(X_1, e_{1.23})}{\sqrt{V(X_1)V(e_{1.23})}} \tag{a}$$

Now $\operatorname{cov}(X_1, e_{1.23}) = E[\{X_1 - E(X_1)\}\{e_{1.23} - E(e_{1.23})\}]$

$$= E[X_1, e_{1.23}]$$

$$= \frac{1}{n}\sum X_1(X_1 - X_{1.23})$$

$$= \frac{1}{n}\sum X_1^2 - \frac{1}{n}\sum X_{1.23}^2$$

$$= \sigma_1^2 - \sigma_{1.23}^2 \tag{b}$$

Also,

$$V(e_{1.23}) = E\{e_{1.23} - E(e_{1.23})\}^2$$

$$= E(e_{1.23})^2$$

$$= \frac{1}{n}\sum (X_1 - X_{1.23})^2$$

$$= \frac{1}{n}\sum X_1^2 - \frac{1}{n}\sum X_{1.23}^2 = \sigma_1^2 - \sigma_{1.23}^2 \tag{c}$$

and

$$V(X_1) = E\{X_1 - E(X_1)\}^2$$

$$= E(X_1)^2$$

$$= \frac{1}{n}\sum(X_1)^2 = \sigma_1^2 \qquad \text{(d)}$$

On substituting (b), (c), and (d) in (a), we get

$$\therefore \qquad R_{1.23} = \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sqrt{\sigma_1^2 . \sigma_1^2 - \sigma_{1.23}^2}}$$

Squaring both sides, we get

$$R_{1.23}^2 = 1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}$$

$$\Rightarrow \qquad 1 - R_{1.23}^2 = \frac{\sigma_{1.23}^2}{\sigma_1^2} = \frac{w}{w_{11}}$$

Where,

$$w = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

$$= \qquad 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}$$

and $\qquad w_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2$

$$\therefore \qquad R_{1.23}^2 = \frac{w_{11} - w}{w_{11}} = \frac{r_{12}^2 + r_{13}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

Hence

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

16.6 LIMITATIONS OF MULTIPLE CORRELATION ANALYSIS:

1) Multiple correlation analysis is based on the assumption that the relationship between the variables is linear. But in practice, most relationships are not linear but follow some other pattern.

2) The effects of independent variables on the dependent variables are separate, distinct and additive.

3) Linear multiple correlation involves a great deal of work relative to the results frequently obtained. When the results are obtained, only well-trained in the method are able to interpret them.

EXAMPLE: In a trivariate distribution,

$$r_{12} = 0.6,\ r_{23} = r_{31} = 0.5$$

Find $R_{1.23}$.

SOLUTION:

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

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

$$= \sqrt{\frac{0.36 + .64 - .768}{0.36}}$$

$$= .803.$$

18.7 PROPERTIES OF MULTIPLE CORRELATION COEFFICIENT:

1) The coefficient of multiple correlation always lies between 0 and 1.

2) If $R_{1.23} = 0$, it means that X_1 is completely uncorrelated with X_2 and X_3.

3) If $R_{1.23} = 1$, this means association is perfect and all the regression residuals are zero.

4) $R_{1.23} > r_{12}, r_{31}$.

16.8 SELF ASSESSMENTS:

1) If $r_{12} = 0.5$, $r_{23} = 0.3$ and $r_{13} = 0.6$, Calculate multiple correlation $R_{2.13}$, $R_{3.12}$.

2) In a trivariate distribution

$$R_{1.23} = \sqrt{\left\{b_{12.3}\, r_{12} \frac{\sigma_2}{\sigma_1} + b_{13.2}\, r_{13} \frac{\sigma_3}{\sigma_1}\right\}}$$

3) (a) Define a multiple correlation for a trivariate distribution. In the usual notations, prove that

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

(b) Show that R is always positive.

4) If $r_{23} = 0$, Prove that $R_{1.23}^2 = r_{12}^2 + r_{31}^2$ and $\sigma_{1.23}^2 = \left(1 - r_{12}^2 - r_{13}^2\right).\ \sigma_1^2$.

5) On the basis of observations made on 35 cotton plants the total correlations of yield of cotton (x_1), number of balls, i.e. seed vessels, (x_2) and height (x_3) are found to be

$r_{12} = 0.863$, $r_{13} = 0.648$ and $r_{23} = 0.709$.

Determine the multiple correlation $R_{1.23}$ and the partial correlation $r_{12.3}$ and $r_{13.2}$ and interpret your results.

****

UNIT - IV **LESSON NO - 17**

PARTIAL CORRELATION COEFFICIENT

STRUCTURE:

17.1 INTRODUCTION:

The correlation and regression coefficients discussed earlier measure the degree and nature of the effect of one variable on another. While it is useful to know how one phenomenon is influenced by another, it is also important to know how one phenomenon is affected by several other variables.

It is often important to measure the correlation between a dependent variable and one particular independent variable when all other variables involved are kept constant i.e. when the effects of all other variables are removed. This can be obtained by calculating coefficient of partial correlation.

For example, if we have three variables–yield of wheat, amount of rainfall and temperature and if we limit our analysis of yield and rainfall to periods when a certain average daily temperature existed or if we treat the problem mathematically in such a way that changes in temperature are allowed for, the problem becomes one of partial correlations.

Thus partial correlation analysis measures the strength of the relationship between Y and one independent variable in such a way that variations in the other independent variables are taken into account.

A partial correlation coefficients is analogous to a partial regression coefficient in that all other factors are "held constant".

The basic distinction between multiple and partial correlation analysis is that whereas in the former we measure the degree of relationship between the variable Y and all the variables X_1, X_2,……..,X_n, taken together, in the latter we measure the degree of relationship between Y and one of the variables X_1, X_2,……..,X_n, with the effect of all the other variables removed.

17.2 OBJECTIVES:

The main objectives of this lesson are:

- to introduce the concept of partial correlation.
- to mention various properties of partial correlation.

17.3 PARTIAL CORRELATION COEFFICIENT:

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

If we denote by $r_{12.3}$ the coefficients of partial correlation between X_1 and X_2 keeping X_3 constant. We find that

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

Similarly,

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

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

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

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

Thus, for three variables X_1, X_2 and X_3 there will be three coefficients of partial correlation each studying the relationship between two variables when the third is held constant.

It should be noted that the value of a partial correlation coefficient is always interpreted via the corresponding coefficient, partial determination, i.e. by squaring the partial correlation coefficient. Thus, if X_1, X_2 and X_3 represent sales, advertisement expenditure and price respectively and we get $r_{12.3}^2 = 0.912$, it means that more than 91 percent of the variation in sales that is not associated with price, is associated with advertisement expenditure.

<u>THEOREM:</u> In a trivariate distribution, prove that Partial correlation between X_1 and X_2 is given by

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

<u>PROOF:</u>

The Partial Correlation between X_1 and X_2 in the trivariate distribution is the correlation between X_1 and X_2 when the influence of X_3 is eliminated. The residual is denoted by $X_{1.3}$,

$$X_{1.3} = X_1 - r_{13}\frac{\sigma_1}{\sigma_3}X_3$$

Similarly,

$$X_{2.3} = X_2 - r_{23}\frac{\sigma_2}{\sigma_3}X_3.$$

The Partial Correlation Coefficients may be denoted by $r_{12.3}$

$$r_{12.3} = \frac{\operatorname{cov}(X_{1.3}, X_{2.3})}{\sqrt{V(X_{1.3})V(X_{2.3})}} \qquad (1)$$

Now,

$$\operatorname{cov}(X_{1.3}, X_{2.3}) = \frac{1}{N}\sum X_{1.3}X_{2.3}$$

$$= \frac{1}{N}\sum\left\{X_1 - r_{13}\frac{\sigma_1}{\sigma_3}X_3\right\}\left(X_2 - r_{23}\frac{\sigma_2}{\sigma_3}X_3\right)$$

$$= r_{12}\sigma_1\sigma_2 - r_{23}\frac{\sigma_2}{\sigma_3}r_{13}\sigma_1\sigma_3 - r_{13}\frac{\sigma_1}{\sigma_3}r_{23}\sigma_2\sigma_3 - r_{13}r_{23}\frac{\sigma_1\sigma_2}{\sigma_3^2}\sigma_3^2$$

$$= \sigma_1\sigma_2(r_{12} - r_{13}r_{23}) \qquad (2)$$

$$V(X_{1.3}) = \frac{1}{N}\sum X_{1.3}^2$$

$$= \frac{1}{N}\sum\left\{X_1 - r_{13}\frac{\sigma_1}{\sigma_3}X_3\right\}^2$$

$$= \sigma_1^2 - 2r_{13}\frac{\sigma_1}{\sigma_3}r_{13}\sigma_1\sigma_3 + r_{13}^2\frac{\sigma_1^2}{\sigma_3^2}\sigma_3^2$$

$$= \sigma_1^2(1 - r_{13}^2) \qquad (3)$$

Similarly,

$$V(X_{2.3}) = \sigma_2^2(1 - r_{23}^2) \qquad (4)$$

On substituting (2), (3) and (4) in (1), we get the desired result i.e.

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

THEOREM: For a trivariate distribution, Show that

$$1-R_{1.23}^{2} = \left(1-r_{12}^{2}\right)\left(1-r_{13.2}^{2}\right)$$

PROOF:

We know that

$$1-R_{1.23}^{2} = \frac{w}{w_{11}} \quad (1)$$

We can easily see that

$$1-r_{13.2}^{2} = 1-\frac{w_{13}^{2}}{w_{11}w_{33}} \quad (2)$$

On dividing (1) by (2), we get

$$\frac{1-R_{1.23}^{2}}{1-r_{13.2}^{2}} = \frac{w\,w_{33}}{w_{11}w_{33}-w_{13}^{2}}$$

$$= \frac{\left(1-r_{12}^{2}\right)\left(1-r_{23}^{2}-r_{31}^{2}-r_{12}^{2}+2r_{12}r_{13}r_{23}\right)}{\left(1-r_{23}^{2}\right)\left(1-r_{12}^{2}\right)-\left(r_{12}r_{23}-r_{13}\right)^{2}}$$

Hence

$$1-R_{1.23}^{2} = \left(1-r_{12}^{2}\right)\left(1-r_{13.2}^{2}\right)$$

This proves the result.

17.4 LIMITATIONS OF PARTIAL CORRELATION COEFFICIENTS:

1) The zero-order correlation must have linear regression.

2) The effects of the independent variables must be additively and not jointly related.

3) Partial analysis possesses the limitations of laborious calculations and difficult interpretation even for statisticians.

17.5 Properties of Partial correlation coefficient:

1) Since each of the correlation coefficient r_{12}, r_{13} and r_{23} lies between $\pm$ 1. So, also

$r_{12.3}$ lies between ± 1.

2) If $r_{12.3} = 0$ we have $r_{12} = r_{13}\, r_{23}$ it means that r_{12} will not be zero if X_3 is correlated with both X_1 and X_3.

3) The expressions for $r_{13.2}$ and $r_{23.1}$ can be similarly obtained as

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{\left(1-r_{12}^2\right)}\sqrt{\left(1-r_{23}^2\right)}} \text{ and } r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{\left(1-r_{12}^2\right)}\sqrt{\left(1-r_{13}^2\right)}}$$

4) Partial correlation coefficient helps in deciding whether to include or not an additional variable in regression analysis.

5) If we assume

$$\Delta = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} \text{ and}$$

Δ_{ij} = Cofactor of the element in the ith row and jth column, then

$$r_{12.3} = \frac{\Delta_{12}}{\sqrt{\Delta_{11}\Delta_{22}}}$$

6) $r_{12.3} = \sqrt{b_{12.3}\, b_{21.3}}$.

EXAMPLE: On the basis of the following information compute:

1) $r_{23.1}$ 2) $r_{13.2}$ and 3) $r_{12.3}$

$r_{12} = 0.70$; $r_{13} = 0.61$; $r_{23} = 0.40$

Solution:

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

Substituting the given values

$$r_{23.1} = \frac{0.4 - 0.7 \times 0.61}{\sqrt{\left(1-(0.7)^2\right)\left(1-(0.61)^2\right)}}$$

$$= - 0.048$$

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

Substituting the given values

$$r_{13.2} = \frac{0.61 - 0.7 \times 0.4}{\sqrt{\left(1-(0.7)^2\right)\left(1-(0.4)^2\right)}}$$

$$= 0.504$$

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

Substituting the given values

$$r_{12.3} = \frac{0.7 - 0.61 \times 0.4}{\sqrt{\left(1-(0.61)^2\right)\left(1-(0.4)^2\right)}}$$

$$= 0.629.$$

Example: Show that the correlation coefficient between the residuals $x_{1.23}$ and $x_{2.13}$ is equal and opposite to that between $x_{1.3}$ and $x_{2.3}$.

Solution: \we have

$$\text{Cov}(x_{1.23},\ x_{2.13}) = \frac{1}{N}\sum\left(x_1 - b_{12.3}x_2 - b_{13.2}x_3\right)x_{2.13}$$

Now, $\sum x_1 x_{2.13} = \sum x_3 x_{2.13} = 0$; $\sum x_2 x_{2.13} = \sum x_{2.13}^2$

$$\therefore \quad \text{Cov}(x_{1.23},\ x_{2.13}) = -\frac{b_{12.3}}{N}\sum x_{2.13}^2$$

$$= -b_{12.3}\sigma^2_{2.13}$$

Hence correlation coefficient between $x_{1.23}$ and $x_{2.13}$ is

$$r\,(x_{1.23}\,,\;x_{2.13}) \;=\; \frac{\operatorname{cov}(x_{1.23}, x_{2.13})}{\sigma_{1.23}\sigma_{2.13}}$$

$$= \frac{-b_{12.3}\sigma_{2.13}}{\sigma_{1.23}}$$

$$= -\frac{\operatorname{cov}(x_{1.3}, x_{2.3})}{\sigma^2_{2.3}}.\frac{\sigma_{2.3}}{\sigma_{1.3}}$$

$$= -\frac{\operatorname{cov}(x_{1.3}, x_{2.3})}{\sigma_{1.3}\sigma_{2.3}}$$

$$= -\;r\,(x_{1.3}\,,\;x_{2.3}).$$

Example: Prove that

$$\sigma^2_{1.23} \;=\; \sigma^2_1\left(1-r^2_{12}\right)\left(1-r^2_{13.2}\right)$$

$$= \sigma^2_{1.2}\left(1-r^2_{13.2}\right)$$

Solution: we know that

$$R^2_{1.23} \;=\; 1-\frac{w}{w_{11}}$$

$$1-R^2_{1.23} \;=\; \frac{\sigma^2_{1.23}}{\sigma^2_1}$$

Now, $\sigma^2_{1.23} = \dfrac{1}{n}\sum X^2_{1.23}$

$\Rightarrow$ $n\sigma^2_{1.23} = \sum X^2_{1.23}$

$=\qquad \sum X_{1.2}^{2} - b_{13.2}\sum X_{1.2}X_{3.2}$

$=\qquad \left[1 - b_{13.2}\frac{\sum X_{1.2}X_{3.2}}{\sum X_{1.2}^{2}}\right]\sum X_{1.2}^{2}$

$=\qquad [1 - b_{13.2}b_{31.2}]n\sigma_{1.2}^{2}$

$\Rightarrow\qquad \sigma_{1.23}^{2} = [1 - r_{13.2}^{2}]\sigma_{1.2}^{2} \qquad (2)$

$\because\qquad b_{13.2}b_{31.2} = r_{13.2}^{2}$

But $\qquad \sigma_{1.2}^{2} = \sigma_{1}^{2}\left(1 - r_{12}^{2}\right) \qquad (3)$

Thus from (2) and (3), we get the required result.

17.6 SELF ASSESSMENT:

1) Explain Partial correlation.

2) Distinguish between multiple and partial correlation.

3) Suppose that $r_{12.34} = 0.5$ and N= 41. Is the partial r significant? Also determine 95% confidence limits.

4) From the data relating to the yield of dry bark (X_1), height (X_2) and girth (X_3) for the 18 cinchona plants, the following correlation coefficient were obtained:

   $r_{12} = 0.77$, $r_{13} = 0.72$ and $r_{23} = 0.52$.

   Find the partial correlation coefficient of $r_{12.3}$.

****

UNIT V **LESSON-18**

SCALES OF MEASUREMENT OF DATA

STRUCTURE:

5.1 INTRODUCTION

In our daily life as we are said to measure when we use yardsticks to determine weight, height, or some other features of a physical object. We also measure when we judge how well we like a song, a painting or the personality of our friends. We, thus, measure physical objects as well as abstract concepts, measurement is a relatively complex and demanding taste, specially so when it concerns qualitative or abstract phenomenon other examples of qualitative characteristic are taste, honesty, intelligence, loyalty etc.

It is easy to assign numbers in respect of properties of some objects, but it is relatively difficult in respect of others, for instance measuring such things as social conformity, intelligence or marital adjustment is much less obvious and requires much closer attention than measuring physical weight, biological age or a person's financial assets. In other words, properties like weight, height etc can be measured directly with some standard unit of measurement, but it is not that easy to measure properties like motivation to succeed, ability to stand stress. For the meaningful assessment of the qualitative characteristics it is essential that they are also measured

5.2 QUANTITATIVE OF QUALITATIVE DATA:

Measurement is defined as a process of associating numbers or symbols to observation obtained in research study. These observation could be Qualitative or Quantitative.

Most of the analysis can be conducted using Quantitative data. For example, mean standard deviation etc. can be computed for Quantitate characteristics. Qualitative characteristics can be counted and cannot be computed. Therefore the researcher must have a clear understanding of the type of characteristic or variable before collecting the data.

The observations on Qualitative variables may also be assigned numbers. For example we can record a person's marital status as 1,2,3 or 4 depending on whether the person is single, married, widowed or divorced. We an as well record 'Yes' or 'No' to a question as 'O' or 'I'. In this artificial or nominal way, categorical data (qualitative or Descriptive) can be made into numerical data and if we thus code the various categories, we refer to the numbers e record as nominal data.

Nominal Data is numerical in name only because they do not share any of the properties of the numbers we deal in ordinary aitthmatic. For instance if we record marital status as 1,2,3 o4 r as stated above, we cannot write 4>2 or 3<4 or we cannot write 3-1 = 4-2 = 4 or4÷2=2.

In those situations when we cannot do anything except setup inequalities, we refer to the data as ordinal data for instance, if one mineral can scratch another, it receives a higher hardness number and on Moh's Scale the number from 1 to 10 are assigned respect to tale, gypsum, calcite, fluorite, apatite, feldspar, quartz, topaz sapphire & diamond with these numbers e can write 5>2 or 6<9 as apatite is harder then gypsum and fledspar is softer than sapphire, bur we can't write for example 10-9 =5-4 because the difference in hardness between diamond and sapphire is actually much greater than that between apatite & fluorite, It would also be meaningless to say that topaz is twice as har as fluorite simply because their respective hardness numbers on Moh's scale are

8 and 4. The gracler than symbol (i.e,>) in connection with ordinal data may be used to designate ' happier than' ' Preferred to' & so on.

When in addition to setting up inequalities we can also form differences, we refer to the data as interval data. Suppose we are given the following temperature readings (in degree Fahrenheit). 58º, 63º, 70º, 95º, 110º, 126º & 135º. In this case we can write 100º>70º or 95º <135º which simply means that 100º is warmer than 70º and that 95º is cooler than 135º. We can also write 95º-70º =135º-110º, since equal temperature differences are equal in the scene that the same amount of heat is required to raise the temperature of an object from 70º to 95º or from 110º to 135º.

When in addition to setting up inequalities and forming differences we can also from quotients (i.e. when we can perform all the customary operations of mathematics) we refer to such data as ratio data. In this sense, ratio data includes all the usual measurement length, height, money weight, volume, area, pressure etc.

The above stated distinction between nominal, ordinal, interval & ratio data is important for the nature of a set of data may suggest the use of particular statistical techniques

18.2 OBJECTIVES:-

In this chapter

1) An understanding of the four level of measurement that can be taken by researchers.

2) The ability to distinguish between comparative & non- comparative measurement scales.

3) The basic too kit that can be used for the purpose of marketing research.

18.3 CLASSIFICATION OF MEASUREMENT SCALES

From what has been stated above, we can write that scales of measurement can be considered in terms of their mathematical properties. The most widely used classification of measurement scales are

a) Nominal Scale

b) Ordinal Scale

c) Intrval Scale

d) Ratio Scale

a) Nominal Scale:- Nominal Scale is simply a system of assigning number symbols to events in order to label them. The usual example of this is the assignment of numbers of basketball players in order to identify them. Such numbes cannot be considered to be associated with an ordered scale for their order is of no consequence; the numbers are just convenient labels for the particulars class of events and as such have no quantitative values. Nominal scale prove convenient ways of keeping trace of people, objects and events. One cannot do much with the number involved.

Nominal scale is the least powerful level of measurement. it indicates no order or distance relationship and has simply differences between things by assigning them to categories. Nominal data is, thus, counted data. Nominal scale are very useful and are widely used in survey and other ex-post-facto research when data is being classified by major subgroups of the pop^n.

Ordinal Scale: The lowest level of the ordered scale that is commonly used is the ordinal scale. The ordinal scale places events in order, but there is no attempt to make the intervals of the scale equal in terms of some rule. Rank order represent the ordinal scales and are frequently used in research relating to qualitative phenomena. A student's rank in his graduation class involves the use of ordinal scale one has to be very careful in making statement about scores based on ordinal scales. For instance, if Ram's position is 40, it can't be said that Ram's position is four times as good as that of Mohan's position. The statements would make

no sense at all. The ordinal scale only permit the ranking of items from highest to lowest. Ordinal scale has no absolute values & the real difference between adjacent ranks may not be equal.

Interval Scale:- Interval sales are numeric scales in which we know not only the order, but also the exact differences between the values. The classic example of an interval scale is celsius temperature because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 & 70 degrees. time is another good example of an interval scale in which the increments are known, consistent and measurable.

Interval Scale are good because the realm of statistical analysis on these data sets open up. For example, central tendency can be measured by mode, median, or mean standard deviation can also be calculated.

Like the others you can remember the key points of an ‘ Interval Scale’ practy lasily. “ Interval” itself means “ Space in between”, which is the important thing to remember interval scale not only tell us about order, but also about the value between each item.

Here’s the problem with interval scale they don’t have “ True Zero”. it does not have the capacity to measure the complete absence a trait or characteristic.

Interval scales provide more powerful measurement than ordinal scale for interval scale also incorporates the concept of equality of interval.

Ratio Scale:- Ratio Scale have an absolute or true zero of measurement. The term “ absolute zero” is not as precise as it as once believed to be. We can conceive of an absolute zero of length & illy e can conceive of an absolute zero of time. For example, the zero point on a centimeter scale indicates the complete absence of length or height. Bit an absolute zero of temperature is theoretically unobtainable & it remains a concept existing only in the scientist’s mind.

Ratio Scale represents the actual amount of variable, measures of physical dimensions such as weight, height distance etc. are examples. generally, all statistical techniques are usable with ratio scale and all manipulations that one can carry out with real numbers can also be carried out with ratio scale values multiplication & division can be used with this scale but not with other scales mentioned above.

18.4 SUMMARY:-

In summary, nominal variables are used to “ name”, or label a series of values ordinal scales provide good information about the order of choices, such as in a customer satisfaction survey. Interval scales give us the order of values & the ability to quantify the difference between each other. Finally, Ratio Scale gives us the ultimate-order, interval values, plus the ability to calculate ratios since a “ True Zero” can be defined.

Measurement is the process of mapping aspects of a domain on to other aspects of a range according to some rules of correspondence.

Different scales are device for different aspects of measurement in statistical research these scales are classified on the basis of

a) Subject Orientation

b) response form

c) Degree of Subjectivity

d) Scale properties

e) No. of dimensions

f) Scale constructions techniques

Spacing of scale construction techniques five main techniques are involved.

a) Arbitrary Approach

b) Consensus Approach

c) Item analysis approach

d) Cumulative Scale

e) Factor Scales.

All the scale falls into these domains and the objective of statistical research are well served.

18.5 SELF ASSESSMENT QUESTIONS

Q1. Discuss the use of scale of measurements of data in statistics?

Q2. Write a short note on

a) Ordinal data

b) Nominal data

c) Ratio data

d) Interval data

UNIT – V LESSON -19

INTRODUCTION OF ATTRIBUTES

STRUCTURE:

19.1 INTRODUCTION:

Normally statistical methods deal with quantitative data only. The quantitative data may arise in two different ways

1. In the first place, the observer may measure the actual magnitudes of some variable character for each of the objects or individuals observed. He may record, for instance, the ages of persons at the time of marriage, heights of students, expenditure of laborers of a factory, amount of agricultural production etc. Data regarding such phenomenon is called statistics of variables.

2. In the second place, the observer may note the presence or absence of some attribute in a series of objects or individuals. Sometimes each observation tells that a particular quality or attribute is present or absent. Observations of this kind are still amenable to statistical treatment, simply by counting or enumerating the number of individuals in which a given quality is present and the number in which it is not present. There are certain phenomena like blindness marriage, deafness etc which cannot be measured. In such cases their presence or absence only can be studied. The quantitative Character in such cases arises solely by counting. Such data are called statistics of attributes.

Literally an attribute means a quality or characteristic. Theory of attributes deals with qualitative characteristics which are not amendable to quantitative measurements and hence need slightly different statistical treatment from that of variables.

For example, according to the attribute “Blindness” the people of a particular city may be classified into two classes.

1. The class of blind people
2. The class of non-blind people

The classification which divides a group into two classes according to one attribute is called classification by dichotomy or simple classification. More than two classes according to one attribute is called manifold classification. For example according to the attribute “Hair colour” the population of a city may be divided into different following classes:

1. Fair haired people
2. Red haired people
3. Brown haired people
4. Black haired people

If several i.e. more than two attributes are noted, the process of classification may be continued indefinitely. Such type of classification may be called classification as a series of dichotomies. For example, consider the two attributes viz “blindness and deafness”. The people of a particular city may be first divided into two classes according to the attribute “blindness and the each of these two classes may further be classified according to the attribute “Deafness” and therefore, ultimately we have following four classes.

1. The class blind and deaf people
2. The class of blind and non-deaf people
3. The Class of non-blind and deaf people
4. The class of non blind and non deaf people

19.2 OBJECTIVES:

The objectives of the present lesson are

- to introduce the concept of attributes
- to aware about the notation and terminology of attributes.
- to know about the combination of attributes

19.3 NOTATION:

Suppose the population is divided into two classes, according to the presence or absence of a single attribute.

The positive class, which denotes the presence of the attribute is generally written in capital Roman letters such as A,B,C, etc. and the negative class, denoting the absence of the attribute is written in corresponding small greek letters such as $\alpha, \beta, \gamma,$ etc. For example if A represents the attribute sickness and B represents blindness, than α and β represent the attributes non-sickness (health) and sight respectively. The two classes viz., A (possession of the attribute) and α (dispossession of the attributes) are said to be complementary classes and the attribute α used in the sense of not-A is often called the complementary attribute of A. similarly β, γ are the complementary attributes of B, C respectively.

19.4 COMBINATION OF ATTRIBUTES:

Combination of attributes is denoted by combination of letters. Thus if 'A' denotes literacy and 'B' criminality, 'AB' will denote literacy and criminality and 'ab' will denote not literacy and not criminality. If we take third attribute C denotes punishment then 'ABC' will denote the individual who is literate and criminal and got the punishment.

'ABg' represent literate and criminal but not got punishment.

'aBC' represent a person who is illiterate and criminal and got the punishment etc.

19.5 DICHOTOMY:

If the universe or population is divided into two sub classes means the population is split into two parts. The half of the members contains in one part and the rest of the members contained in second. In other words, the classes are complementary. There are no more elements with respect to each of the attributes A, B, C etc. thus the division or classification is said to be ‘dichotomous classification. The classification is termed manifold if each class is further sub divided.

****

UNIT - V **LESSON - 20**

CLASS FREQUENCY, ULTIMATE CLASS AND CONTINGENCY TABLE

STRUCTURE:

20.1 INTRODUCTION:

In the previous lesson, we learnt about the meaning of attributes and various notations and terminology but in the present lesson we will be know about the class frequency and ultimate class frequency. What is contingency table and how to complete the table? The number of observation or persons in different classes are called "class frequencies". The class frequencies are denoted by the class notation within bracket, like (A), (B), (α), (β), (AB), (Aβ), ($\alpha\beta$), (ABC) etc

The order of a class depends upon the number of attributes specified. A class having one attribute is known as the class of the first order, a class having two attributes is known as second order class etc.

20.2 CLASS FREQUENCY:

The number of individuals or units belonging to a class is known as its frequency. For convenience and clear understanding, the frequency of a class is denoted by the letters is a parenthesis representing that cell. For example, the frequency of the class AB is denoted by (AB).

(A): denotes the number of items possessing A attribute

(α): denotes the number of items not possessing A attribute

(B): denotes the number of items possessing B attribute

(β): denotes the number of items not possessing B attribute

(AB): denotes the number of items possessing A & B attributes

(Aβ): denotes the number of items possessing attribute A but not possessing B attribute

($\alpha\beta$): denotes the number of items not possessing A and B attributes

ULTIMATE CLASS FREQUENCY: The class of highest order is called ultimate class and its frequency is called the ultimate class frequency. Thus in case of (n) attributes, the ultimate class frequencies will be the frequencies of nth order. For example, the class frequencies (ABC), (ABγ), (AβC), (A$\beta\gamma$), (αBC), ($\alpha\beta$C), (αBγ) are the ultimate class frequencies fo three attributes A, B, and C.

NUMBER OF CLASSES: The total number of classes formed on the basis of the number of attributes studied can be known by 3^n formula. Where 'n' stands for the number of attributes studied. If only one attribute is studied, then there will be 3^1 or only three classes (N,A, α). If three attributes are studied, then there will be 33=27 classes in all (N, A, B, C, α, β, γ , AB, AC, BC, Aβ, Aγ, Bγ, $\alpha\beta$, αC, βC, $\alpha\beta$, $\beta\gamma$, abc, ABγ, AβC, A$\beta\gamma$, ABC, αBγ, $\alpha\beta$C, $\alpha\beta\gamma$).

The attributes denoted by capital letters are called positive attributes such as A, AB, BC, ABC etc. the attribute denoted by Greek letters are called negative attributes such as (α, $\alpha\beta$, $\beta\gamma$, $\alpha\beta\gamma$,) etc. classes having a combination of both positive and negative attributes like Aβ, αB, αBγ etc are called pairs of contrary classes.

For three attributes A, B and C, the combinations of attributes belonging to different classes can be displayed in the following manner.

N

(A) (α)

(AB) (Aβ) (αB) (αβ)

(ABC) (ABγ) (AβC) (Aβγ) (αBC) (αBγ) (αβC) (αβγ)

Similar relations can be given N, B, β and N, C, γ.

20.3 ORDER OF THE CLASSES: - The total number of observations in the universe or population is by N and is called the class of the order zero. If two attributes are studied say A,B then A,B, α,β, would be classes of the first order and AB, Aβ, αB, αβ would be the classes of ultimate order .

The total number of classes of ultimate order is determined by the formula 2^n, where 'n' stands for the number of attributes studied. If two attributes are studied the number of classes of ultimate order would by $2^2 = 4$. In case of three attributes, then there would be $2^3 = 8$ classes of ultimate order.

This is illustrated below

		No. of Classes
Frequency of the zero order	N	1
Frequencies of the first order -	(A), (B), (C)	
	(α), (β), (γ)	6
Frequencies of the second order-	(AB), (AC), (BC)	
	(ABβ), (Aγ), (Bγ)	
	(αB), (αC), (βC)	
	(αβ), (αγ), (βγ)	12
Frequencies of the third order -	(ABC), (αBC), (AβC), (Aβγ)	
	(ABγ), (αBγ), (αβC), (αβγ)	8
	Total	$3^3 = 27$

Relation between Class Frequencies: All the class frequencies of various orders are not independent of each other and any class frequency can always be expressed in terms of class frequencies of highest order. Thus

$N = (A) + (\alpha) = (B) + (\beta) = (C) + (\gamma)$ etc

Also, since each of these A’^{S} or α‘^{s} can either be B’^{s} or β’^{s}, we have

$(A) = (AB) + (A\beta)$ and $(\alpha) = (\alpha B) + (\alpha\beta)$

Similarly $(B) = (AB) + (\alpha B)$ and $(\beta) = (A\beta) + (\alpha\beta)$

$(AB) = (ABC) + (AB\gamma)$ and $(A\beta) = (A\beta C) + (A\alpha\gamma)$

$(\alpha B) = (\alpha BC) + (\alpha B\gamma)$ and $(\alpha\beta) = (\alpha\beta C)+(\alpha\beta\gamma)$ and so on

$(A) = (AB) + (A\beta) = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma)$

$(\beta) = (A\beta) + (\alpha\beta) = (A\beta C) + (A\alpha\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$, etc

The frequencies can also be known with the help of a nine square table

	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	($\alpha\beta$)	(β)
Total	(A)	(α)	N

Vertical totals:

$$(AB) + (A\beta) = (A)$$
$$(\alpha B) + (\alpha\beta) = (\alpha)$$
$$(B) + (\beta) = N$$

Horizontal totals

$$(AB) + (\alpha B) = (B)$$
$$(A\beta) + (\alpha\beta) = (\beta)$$
$$(A) + (\alpha) = N$$

EXAMPLE: From the following ultimate class frequencies, find the frequencies of the positive and negative classes and the total number of observations

$(AB) = 100$ $(\alpha B) = 80$

$(A\beta) = 50$ $(\alpha\beta) = 40$

SOLUTION: We have to find out

N, (A), (B), (α) and (β):

This can be found with the help of nine square table

	A	α	Total
B	(AB) 100	(αB) 80	(B) 180
β	$(A\beta)$ 50	$(\alpha\beta)$ 40	(β) 90
Total	(A) 150	(α) 120	N 270

$$N = (A) + (\alpha) = (AB) + (\alpha B) + (A\beta) + (\alpha\beta)$$

$$= 100+80+50+40$$

$$= 270$$

$$(A) = (AB) + (A\beta) = 100+50 = 150$$

$$(B) = (AB) + (\alpha B) = 100 + 80 = 180$$

$$(\alpha) = (\alpha B) + (\alpha\beta) = 80 + 40 = 120$$

$$(\beta) = (A\beta) + (\alpha\beta) = 50 + 40 = 90$$

EXAMPLE: Given the following frequencies of the positive classes. Find the frequencies of the ultimate classes:

$(A) = 160$, $(B) = 200$, $(AB) = 140$, $N = 500$

SOLUTION: We have to find out $(A\beta)$, (αB) and $(\alpha\beta)$

$(A\beta) = (A)-(AB) = 160-140 = 20$

$(\alpha B) = (B)-(AB) = 200-140 = 60$

$$(\alpha\beta) = (\alpha)-(\alpha B) = [N-(A)] - [(B)-(AB)]$$
$$= N- (A) - (B) + (AB)$$
$$= 500-160-200+140$$
$$= 280$$

CLASS SYMBOLS AS OPERATORS :

Let us write symbolically

$$A.N. = (A)$$

Which means that the operation of dichotomizing N according to A gives the class frequency equal to (A)

Similarly, we write

$$\alpha.N = (\alpha)$$

Adding, we get $A.N. + \alpha.N = (A) + (\alpha)$

$$\Rightarrow (A + \alpha).N = N$$
$$\Rightarrow A + \alpha = 1$$

Thus in symbolic expression we can replace A by $(1-\alpha)$ and α by (1-A)

Similarly B can be replaced by $(1-\beta)$ and β by (1-B) and so on

Dichotomizing (B) according to A, let us write

$$A. (B) = (AB)$$

Similarly B. (A) = (BA)

A (B) = B.(A) = (AB) = A.B.N which amounts to dichotomizing N according to AB.

For example $(\alpha\beta) = \alpha\beta.\ N = (1-A)\ (1-B).\ N = N-AN-BN+AB.N$

$$= N-(A)-(B) + (AB)$$

$$(\alpha\beta\gamma) = \alpha\beta\gamma\ .N = (1-A)\ (1-B)\ (1-C).N$$
$$= N-AN-BN-CN+ABN+CAN+BCN-ABCN$$

$= N-(A)-(B)-(C) + (AB) + (AC) + (BC)-(ABC)$

$(AB\gamma) = AB\gamma.N = AB\,(1-C)\,N = ABN-ABCN = (AB)-(ABC)$

$(\alpha BC) = \alpha BC.N = (1-A)\,BCN = BCN-ABCN= (BC)-(ABC).$

EXAMPLE: Obtain all the ultimate class frequencies from the given data

N=23713, (A) =1618, (B) =2015, (C) = 770

(AB)=587, (BC) = 428, (AC) =335, and (ABC) =156.

SOLUTION: For three attributes A, B and C the number of ultimate class frequencies in 23 =8. So one of them (ABC) is given. The remaining are

$(AB\gamma)$, $(A\beta C)$, (αBC), $(A\beta\gamma)$, $(\alpha B\gamma)$, $(\alpha\beta C)$, $(\alpha\beta\gamma)$

Now $(AB\gamma) = (AB)-(ABC) = 587-156 = 431$

$(A\beta C) = (AC)-(ABC) = 335-156 = 179$

$(\alpha BC) = (BC)-(ABC) = 428-156 = 272$

$(A\beta\gamma) = A\beta\gamma.\ N = A\,(1-B)\,(1-C)\,N = (A) - (AB)-(AC) + (ABC)$

$= 1618-587-335+156 =852$

$(\alpha B\gamma) = (B)-(AB)-(BC) + (ABC)$

$= 2015-587-428+156 =1156$

$(\alpha\gamma C) = (C) - (AC)-(BC) + (ABC)$

$= 770-335-428+156=163$

$(\alpha\beta\gamma) = N - (A) - (B) - (C) + (AB) + (BC) + (AC) - (ABC)$

$= 237.3 – 1618 – 2015 –770 + 587 + 428 + 335 –156$

$= 20504.$

20.5 Summary: In this lesson we have discussed

ATTRIBUTES: Qualitative characteristic are called as attributes

POSITIVE ATTRIBUTES: The presence of attributes is called positive attributes and denoted by capital letter A, B, C etc

NEGATIVE ATTRIBUTES: The absence of attributes is called negative attribute and denoted by Greek letter α , β , γ etc

CLASS FREQUENCY: - The number of observation assigned to any class is called frequency. It is written by enclosing the class symbols in brackets such as (A), (AB) etc.

ULTIMATE CLASS FREQUENCY: - The class specified by n attributes i.e. those of the highest order are called ultimate class frequency .

20.6 SELF ASSESSMENTS:

1. With three attributes A, B, and C, write down

   a) Number of positive class frequencies

   b) Number of ultimate class frequencies.

   c) Number of all the class frequencies.

   d) All the class frequencies in symbols.

2. Given that (AB) = 150, (Aβ) = 230, (αB) = 260, ($\alpha\beta$) = 2340. Find all other frequencies and the value of N.

3. Given following ultimate class frequencies. Find the frequencies of positive classes.

   (ABC) = 149, (ABγ) = 738, (AβC) = 225, (A$\beta\gamma$) = 1196,

   (αBC) = 204, (αBγ) = 1762, ($\alpha\beta$C) = 171 and ($\alpha\beta\gamma$) = 21842

****

UNIT - V **LESSON-21**

CONSISTENCY OF DATA, CONDITIONS OF CONSISTENCY OF DATA

STRUCTURE:

21.1 INTRODUCTION:

In the lesson-26 of attributes, we learnt about class frequency, ultimate class and contingency table.

In the present lesson we will study about consistency of data and its conditions.

The statistics of attributes are obtained by counting and therefore no class frequency can be negative. If any class frequency has a negative value, data are said to be inconsistent. Such inconsistency may be due to wrong counting or inaccurate or subtractions or may be due to misprints.

21.2 CONSISTENCY OF DATA :

The data are said to be consistency if any of the ultimate frequency associated with the same population is not negative. Hence, the data are said to be inconsistent if they lead to any frequency obtained by the relations among frequencies as negative. Also no higher order class can have a greater frequency than the lower order class frequency.

It should be, however, remembered that consistency of data is no proof of accurate counting or printing, though the inconsistency of data is a sure proof of data being incorrect. To find out consistency of data, ultimate class frequencies should be found out, because of there is any inconsistency, one or more ultimate class frequencies will be negative.

For instance if (A) = 40, (AB) =42, So (A) and (AB) are inconsistent as (AB) cannot be greater than (A) if they are observed from the same population.

On the basis of there consideration, it can be stated that "The necessary and sufficient condition for consistency of a set of independent class frequencies is that no class frequency must be negative".

21.3 CONDITIONS FOR CONSISTENCY OF DATA :

When the data pertaining to one or single attribute A, we have conditions of consistency are as follows

1) $(A) \geq 0$

2) $(\alpha) \geq 0 \Rightarrow \alpha N \geq 0$

$\Rightarrow \quad (1\text{-}A)\ N \geq 0$

$\Rightarrow \quad N\text{-}AN \geq 0$

$\Rightarrow \quad N\text{-}(A) \geq 0 \quad$ or $(A) \leq N$

When the data pertaining to two attributes viz A and B, the conditions of consistency are:

i) $(AB) \geq 0$

ii) $(A\beta) \geq 0 \quad \Rightarrow \quad (AB) \leq (A)$

iii) $(\alpha B) \geq 0 \quad \Rightarrow \quad (AB) \leq (B)$

iv) $(\alpha\beta) \geq 0 \quad \Rightarrow \quad \alpha\beta N \geq 0 = (1\text{-}A)\ (1\text{-}B)\ N \geq 0$

$\Rightarrow \quad N\text{-}AN\text{-}BN\text{+}ABN \geq 0$

$\Rightarrow \quad N\text{-}(A)\text{-}(B) + (AB) \geq 0$

$\Rightarrow \quad (AB) \geq (A) + (B)\text{-}N$

When the data pertaining to three attributes viz A, B, and C then condition of consistency are

1) $(ABC) \geq 0$

2) $(AB\gamma) \geq 0 \quad \Rightarrow (ABC) \leq (AB)$

3) $(A\beta C) \geq 0 \quad \Rightarrow (ABC) \leq (AC)$

4) $(\alpha BC) \geq 0 \quad \Rightarrow (ABC) \leq (BC)$

5) $(A\beta\gamma) \geq 0 \quad \Rightarrow (ABC) \geq (AB) + (AC)-(A)$

6) $(\alpha B\gamma) \geq 0 \quad \Rightarrow (ABC) \geq (AB) + (BC) - (B)$

7) $(\alpha\beta C) \geq 0 \quad \Rightarrow (ABC) \geq (AC) + (BC) - (C)$

8) $(\alpha\beta\gamma) \geq 0 \quad \Rightarrow (ABC) \leq N - (A)-(B)-(C) + (AB) + (BC) + (AC)$

Since $(ABC) > 0$ and $(\alpha\beta\gamma) \geq 0$, we have

$$(AB) + (BC) + (AC) \geq (A) + (B) + (C) - N$$

Similarly (ii) and (vii) $\Rightarrow (AC) + (BC) - (AB) \leq (C)$

(iii) and (vi) $\Rightarrow (AB) + (BC) - (AC) \leq (B)$

(iv) and (v) $\Rightarrow (AB) + (AC) - (BC) \leq (A)$

EXAMPLE: consider the following data for two attributes A and B and test its consistency

$(A)= 80$ $(\beta) = 280$ $(AB) = 50$ $N=400$

SOLUTION: To test the consistency, we have to prepare the following table using the relations between class frequencies and ultimate class frequencies.

	A	α	Total
B	(AB) 50	(αB) 70	(B) 120
β	Aβ	(αβ) 250	(B) 280
Total	(A) 80	(α) 320	N 400

$(B) = N-(\beta) = 400-280 =120$

$(\alpha B) = (B) - (AB)$ $120-50 =70$

$(\alpha) = N-(A) = 400-80=320$

$(\alpha\beta) = (\alpha) - (\alpha B) = 320-70 = 250$

$(A\beta) = (A)-(AB) = 80-50 = 30$

Since all the ultimate class frequencies are positive, so the data is consistent.

EXAMPLE: Consider the following data for two attributes A and B and test its consistency

$N = 400\ (A) = 80\ (B) = 120\ (AB) = 130\ (\gamma\gamma) = 330$

SOLUTION: To test the consistency, we have to prepare the following table by using the relation between class frequencies and ultimate class frequencies.

	A	a	Total
B	(AB) 130	(aB) -10	(B) 120
β	$(A\beta)$ -50	$(\alpha\beta)$ 330	(β) 280
Total	(A) 80	(β) 320	N 400

$(\alpha B) = (B) - (AB) = 120-130 = -10$

$(A\beta) = (A) - (AB) = 80-130 = -50$

$(\alpha) = N- (A) = 400 - 80 = 320$

$(\alpha) = N - (B) = 400-120 = 280$

in the above table, we note that the ultimate frequencies of (αB) and $(A\beta)$ are negative. Hence the given data are inconsistent.

EXAMPLE: A market investigator returns the following data of 1000 people consulted, 811 likes chocolates, 752 liked toffee and 418 liked boiled sweets, 570 liked chocolates and toffee, 356 liked chocolates and boiled sweets and 348 like toffee and boiled sweets, 297 liked all three. Show that this information as stands must be incorrect.

SOLUTION: Let A be denoted for liking of chocolates, B be denoted for liking of toffee, C be denoted for liking of boiled sweets, then

$N=1000,\quad (A)=811,\quad (B)=752,\quad (C) = 418,$

$(AB) = 570,\qquad (AC) = 356,\quad (BC) = 348,\ (ABC) = 297$

Let us find out from the given data, whether any of the ultimate class frequency is negative

$(\alpha BC) = (BC)-(ABC) = 348-297 = 51$

$(AB\gamma) = (AB) - (ABC) = 570-297 = 273$

$(A\gamma C) = (AC) - (ABC) = 356-297 = 59$

$(A\beta\gamma) = (A\beta) - (A\beta C) = (A) - (AB) - (A\beta C)$

$= 811-570-59$

$= 182$

$(\alpha\beta C) = (C) - (BC) - (A\beta C) = 418-348-59 = 11$

$(\alpha B\gamma) = (B) - (AB) - (\alpha BC) = 752-570-51 = 131$

$(\alpha\beta\gamma) = N - (A) - (B) - (C) + (AB) + (BC) + (AC) - (ABC)$

$= 1000-811-752-418+570+348+356-297$

$= -4$

Since $(\alpha\ \beta\ \gamma)$ is negative. Hence data are inconsistent.

PROBLEM: Given that $(A) = (\alpha) = (B) = \beta = N/2$

Prove that $(AB) = (\alpha\ \beta)$ and $(A\ \beta) = (\alpha B)$

SOLUTION: $(A) = (AB) + (A\ \beta)$

$(B) = (AB) + (\alpha B)$

As $(A) = (B)$

Hence $(AB) + (A\beta) = (AB) + (\alpha B)$

$\therefore \quad (A\beta) = (\alpha B)$

$(A) = (AB) + (A\beta)$

$(\alpha) = (\alpha B) + (\alpha\ \beta)$

As $(A) = (\alpha) \Rightarrow (AB) + (A\beta) = (\alpha B) + (\alpha\beta)$

But $(A\ \beta) = (\alpha B)$

Thus $(AB) + (\alpha B) = (\alpha B) + (\alpha\beta)$

$\therefore \quad (AB) = (\alpha\ \beta)$ Proved.

PROBLEM: Given that

$(A) = (\alpha) = (B) = (\ \beta) = (C) = (\gamma) = N/2$

and $(ABC) = (\alpha\ \beta\ \gamma)$

Prove that i) $2(ABC) = (AB) + (BC) + (AC) - N/2$

ii) $(\alpha\ \beta\ \gamma) = N/2 - 1/2\ \{(A\ \beta) + (B\ \gamma) + (\alpha C)$

SOLUTION: (i) $(\alpha\ \beta\ \gamma) = (AB) + (AC) + (BC) - (A) - (B) - (C) + N - (ABC)$

Or $(ABC) = (AB) + (AC) + (BC) - (A) - (B) - (C) + N - (ABC)$

$\because \quad (ABC) = (\alpha\ \beta\ \gamma)$

$\Rightarrow \quad 2\ (ABC) = (AB) + (AC) + (BC) - (A) - (B) - (C) + (N)$

$\Rightarrow \quad 2(ABC) = (AB) + (AC) + (BC) - N/2 - N/2 - N/2 + N$

$\Rightarrow \quad 2\ (ABC) = (AB) + (AC) + (BC) - N/2$ Proved.

(ii) $(\alpha\beta\gamma) = (AB) + (AC) + (BC) - (A) - (B) - (C) + N - (ABC)$

$(\alpha\beta\gamma) + (ABC) = (AB) + (AC) + (BC) - (A) - (B) - (C) + N$

$2(\alpha\beta\gamma) = \{(A) - (A\beta)\} + \{(C) - (\alpha C)\} + \{(B) - (B\gamma)\} - (A)-(B)-(C) + N$

$2(\alpha\beta\gamma) = -\ (A\beta)-(\alpha C)-(B\gamma) + N$

$(\alpha\beta\gamma) = N/2-1/2\{(A\beta) + (\alpha C) + (B\gamma)\}$. Proved.

PROBLEM: Show that if

$$\frac{(A)}{N} = X,\ \frac{(B)}{N} = 2X,\ \frac{(C)}{N} = 3X$$

and

$$\frac{(AB)}{N} = \frac{(BC)}{N} = \frac{(CA)}{N} = Y$$

Then the value of neither X nor Y can exceed ¼ .

SOLUTION: Conditions of consistency are:

$(AB) \leq (A)$

$\Rightarrow \quad NY \leq NX$

$\Rightarrow \quad Y \leq X \qquad (1)$

Also $(BC) \geq (B) + (C) - (N)$, dividing b/s by N, we get

$$\frac{(BC)}{N} \geq \frac{(B)}{N} + \frac{(C)}{N} - 1$$

$\Rightarrow \quad Y \geq 2X + 3X - 1$

$\Rightarrow \quad Y \geq 5X - 1$

$5X - 1 \leq Y \qquad (2)$

From (1) and (2) we get

$5X-1 \leq X$

$\Rightarrow \quad 5X-X \leq 1$

$\Rightarrow \quad 4X \leq 1$

$\Rightarrow \quad X \leq \frac{1}{4} \qquad (3)$

From equation (1) and (3), we conclude that

$Y \leq X \leq \frac{1}{4}$. Hence proved.

21.4 <u>SUMMARY</u>:

In this lesson we read about consistency and inconsistency of the data :

Consistency means all the class frequencies are positive when they observed from the same population

Inconsistency means a set of class frequencies which do not conform to the same population and provide contradictory statement. In other words if any class frequencies are negative then the data are inconsistent. The various conditions for the consistency are:

i) $(A) \geq 0$

ii) $(\alpha) \geq 0 \qquad (A) \leq N$

iii) $(AB) \geq 0$

iv) $(\alpha\beta) \geq 0 \qquad \Rightarrow (AB) \geq (A) + (B) - N$

v) $(ABC) \geq 0$

vi) $(\alpha\beta\gamma) \geq 0$

$(ABC) \leq N-(A)-(B)-(C) + (AB) + (BC) + (AC)$, etc.

21.5 SELF ASSESSMENTS:

1. What do you understand by consistency of given data? How do you check it?

2. From the following three cases find out whether the data are consistent are not.

   1) (A) = 100 (B) =150 (AB) = 60 N=500

   2) (A) = 100 (B) =150 (AB) = 140 N=500

   3) (A) = 600 (B) = 500 (AB) = 50 N= 1000

3. A Student reported the results of a survey in the following manner, in terms of the usual notations:

   N=1000, (A) = 525 (B) = 312 (C)= 470 (AB) = 42 (BC) = 86 (AC) = 147 and (ABC) = 25

4. Show that for n attributes A_1, A_2 - - - - - - -A_n

   ($A_{1,}$ A2.................A_n) ≥ (A_1) + (A_2) +...............+ (A_n)-(n-1) N

Where N is the total number of observations.

****

Unit - V **Lesson 22**

Association of attributes and independence

STRUCTURE

22.1 INTRODUCTION:

In the present lesson, we study about the association of attributes which measure the relationship between two such phenomena whose size cannot be measured and where we can only find out the presence or absence of an attribute.

In correlation analysis we study the relationship between two variables whose values can be measured. Similary, in case of association we study the relationship between two atrributes.

The study of association can be done by any one of the following methods:

i) Comparison of actual and observed frequencies.

ii) Comparison of various proportions and products.

iii) Comparison of Yule's coefficient of Association.

iv) Calculation of coefficient of contingency.

22.2 OBJECTIVES:

After reading this lesson you should be able to:

i) Understand the concept of independence of attributes.

ii) Understand various forms of independence

iii) Know the concepts of association and its various conditions.

22.3 ASSOCIATION OF ATTRIBUTES:

Statistical data are generally of two types. One that can be measured quantitatively, e.g., height, weight etc. and other which cannot be measured in figures. e.g. deafness, dumbness etc. It is desired to investigate the relationship between the data of a descriptive character- known association attributes, the method of association is resorted to.

The word 'Association' has a technical meaning in Statistics. Two attributes are said to be associated, if they appear together in a larger number of classes than is to be expected if they are independent. In other words, two attributes A and B are said to be associated if they are not independent but are related in some way or the other.

22.4 TYPES OF ASSOCIATION:

There can be three types of association between the attributes.

i) **POSITIVE ASSOCIATION:** when two attributes are present or absent together in the data, they are said to be positively associated. In such cases, the actual frequency is more than the expected frequency . Such association is found between literacy and employment, smoking and cancer, vaccination and immunity from a disease etc.

ii) **NEGATIVE ASSOCIATION:** When the presence of attributes causes absence of another attributes, they are said to have negative association between them. In such cases, the actual frequency is less than the expected frequency.

iii) **INDEPENDENCE:** When two attributes do not have any tendency to be present together, or one's presences do not cause absence of the other attributes, two attributes are regarded as independent. In such a situation the actual frequency is equal to the expected frequency.

22.5 INDEPENDENCE OF ATTRIBUTES:

Two attributes are said to be independent if there exists no relationship of any kind between them. If A and B are independent, we expect:

i) The same proportion of A's amongst B's as amongst β's.

ii) The proportion of B's amongst A's is same as that amongst the α's.

Mathematically, it can be expressed as,

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \qquad (1)$$

$$\Rightarrow \quad 1-\frac{(AB)}{(B)} = 1-\frac{(A\beta)}{(\beta)}$$

$$\Rightarrow \quad \frac{(B)-(AB)}{(B)} = \frac{(\beta)-(A\beta)}{(\beta)}$$

$$\Rightarrow \quad \frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}$$

Similarly, (ii) gives

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$$

$$\Rightarrow \quad 1-\frac{(AB)}{(A)} = 1-\frac{(\alpha B)}{(\beta)}$$

$$\Rightarrow \quad \frac{(A)-(AB)}{(A)} = \frac{(\alpha)-(\alpha B)}{(\beta)}$$

$$\Rightarrow \qquad \frac{(A\beta)}{(A)} = \frac{(\alpha\beta)}{(\beta)}$$

Again from (i) and (ii), we have

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$$

$$\Rightarrow \qquad \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N}$$

and

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$$

$$\Rightarrow \qquad \frac{(AB)+(\alpha B)}{(A)+(\alpha)}$$

$$= \qquad \frac{(B)}{N} = \frac{(B)-(AB)}{N-(A)} = \frac{(\alpha B)}{(\alpha)}$$

Since we know that

$$\frac{(AB)}{(A)} = \frac{(B)}{N}$$

$$\Rightarrow \qquad (AB) = \frac{(A).(B)}{N} \qquad (2)$$

Dividing both sides by N, we get

$$\frac{(AB)}{N} = \frac{(A)}{N} . \frac{(B)}{N}$$

Thus equation (2) leads to the rule that if two attributes A and B are independent, the frequency of AB in the population under consideration is equal to the product of frequencies of A and B divided by the total frequencies.

22.6 METHOD OF COMPARISON OF OBSERVED AND EXPECTED FREQUENCIES:

When association between two attributes is found by appling this method , the actual frequency is compared with the expected frequency . Thus, in order to find out association between the two attributes, it becomes necessary to find out the expected number of their simultaneous occurrence.

Therefore, the expected frequency of a particular event is:

$$\frac{\text{Number of favourable cases}}{\text{Total number of cases}} \times \text{number of observations}$$

Thus, if a coin is tossed the probability that it will fall head upward in the 100 such tosses is. On the basis of this theory the probability of (A) is $\frac{(A)}{N}$ and that of (B) is $\frac{(B)}{N}$.

The combined probability of (A) and (B) is $\frac{(A)}{N} \times \frac{(B)}{N}$.

The expectation of (A) and (B) jointly is $\frac{(A)}{N} \times \frac{(B)}{N}$. N or $\frac{(A).(B)}{N}$

$$\Rightarrow \quad E\,(AB) = \frac{(A).(B)}{N}$$

Similarly, the expectation of $(\alpha\beta) = \frac{(\alpha).(\beta)}{N}$

In the same way expected frequencies can be found out for others also.

The association is positive if E (AB) > $\frac{(A).(B)}{N}$

The association is negative if E (AB) < $\frac{(A).(B)}{N}$

The association is independent if E (AB) = $\frac{(A).(B)}{N}$.

<u>EXAMPLE</u>: Show that A and B are independent, positively associated or negatively associated in the following cases:

1) N = 300, (A) = 48, (B) = 100 and (AB) = 16

2) N = 120, (A) = 24, (B) = 100 and (AB) = 52

3) N= 400, (A) = 150, (B) = 200 and (AB) = 50.

<u>SOLUTION</u>:

1) The expected frequency of (AB) is $\frac{(A).(B)}{N} = \frac{48 \times 100}{300} = 16$

This is equal to the actual frequency of (AB). Hence A and B are independent.

2) The expected frequency of (AB) is $\frac{(A).(B)}{N} = \frac{24 \times 100}{120} = 20$

The actual frequency of (AB) is 52, which is greater than the expected frequency of (AB). Hence A and B are positively associated.

iii) The expected frequency of (AB) is $\frac{(A).(B)}{N} = \frac{150 \times 200}{400} = 75$

The actual frequency of (AB) is 50, which is less than the expected frequency of (AB). Hence A and B are negatively associated.

22.7 ALTERNATIVE METHOD TO FIND ASSOCIATION, IF ULTIMATE CLASS FREQUENCIES ARE GIVEN:

If all the ultimate class frequencies are given, the association can be found out by the following formulae:

$(\text{AB})\times(\alpha\beta) = (\text{A}\beta)\times(\alpha\text{B})$, No Association.

$(\text{AB})\times(\alpha\beta) > (\text{A}\beta)\times(\alpha\text{B})$, Positive Association.

$(\text{AB})\times(\alpha\beta) < (\text{A}\beta)\times(\alpha\text{B})$, Negative Association.

<u>EXAMPLE</u>: In an anti- malaria campaign in a certain area quinine was administrated to 812 persons out of a total population of 3248.

The number of fever cases is shown below:

Treatment	Fever	No Fever
Quinine	20	792
No- quinine	220	2216

Discuss the usefulness of quinine in checking malaria.

<u>SOLUTION</u>: Denoting A for quinine treatment

α for the No quinine

B for attack of fever

β for no attack of fever. Then

	B		β		Total	
A	(AB)	20	(Aβ)	792	(A)	812
α	(αB)	220	($\alpha\beta$)	2216	(α)	2436
Total	(B)	240	(β)	3008	N	3248

To test association, $(\text{AB}) \times (\alpha\beta) = (\text{A}\beta) \times (\alpha B)$

$\Rightarrow \quad 20 \times 2216 = 220 \times 792$

$\Rightarrow \quad 44320 = 174240$

Since 44320 < 174240, thus there is negative association between A and B. It means that the quinine treatment is not useful in checking malaria.

22.8 YULE'S COEFFICIENT OF ASSOCIATION:

The above discussed method of finding association, between atrributes is simple but it only gives a rough idea about their association. The degree of association cannot be found out. To know the extent of association between two attributes, Prof. Yule has given a formula. In this formula the coefficient of association is denoted by 'Q'. The formula is

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

The coefficient of association 'Q' varies between 1. And it is interpretated like coefficient of correlation. If the result is 0, there is no association between two attributes, when the result is + 1, there is perfect positive association and in case of - 1, there is perfect negative association.

EXAMPLE: In an assortative study, to find whether tall husbands tend to marry tall wives, the following information about the wives of 125 tall and 125 short statured husbands was published,.

	Tall husbands	Short Husbands
Tall wives	56	13
Short wives	11	45

Find the coefficient of association between the stature of wives and husbands.

SOLUTION: Let A denote tall husbands

α denote short husbands

B denote the tall wives

β denote the short wives.

Thus, the given combinations are

$(AB) = 56, (A\beta) = 11, (\alpha B) = 13, (\alpha\beta) = 45$

Therefore the coefficient of association $(Q) = \dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

$$\Rightarrow \qquad Q = \frac{56\times 45 - 11\times 13}{56\times 45 + 11\times 13} = 0.892$$

It shows that a very high degree of positive association between statures of husbands and wives.

22.9 SUMMARY:

In this lesson we have studied about the independence of attributes, association of attributes andcoefficient of association,

i) Independent if $(AB) = \dfrac{(A).(B)}{N}$

ii) Positively associated if $(AB) > \dfrac{(A).(B)}{N}$

iii) Negatively associated if $(AB) < \dfrac{(A).(B)}{N}$.

iv) Yule's coefficient of association is $Q = \dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$.

22.10 SELF ASSESSMENTS:

1) Find if A and B are independent, positively associated or negatively associated, in each of the following cases:

i) N = 1000, (A) = 470, (B) = 620 and (AB) = 320

ii) (A) = 490, (AB) = 294, (α) = 570 and (α B) = 380

iii) (AB) = 256, (α B) = 768, (A β) = 48 and ($\alpha\beta$) = 144.

2) From the following data, discuss if the colour of son's eyes is associated with that of father's.

		Eye color in Son	
		Not light	Light
Eye colour in Father	Not light	230	148
	Light	151	471

3) What is association of attributes? Write a note on the strength of association and how it is measured?

4) The following data relate to literacy and unemployment in a group of 500 persons. You are required to calculate Yule's coefficient of association between literacy and unemployment and interpret it.

Illiterate unemployed	220
Literate employed	20
Illiterate employed	180

****