STATISTICAL METHODS & INFERENCES
(DSSTT21/DBSTT21)
(BSC, BA STATISTICS -II)

ACHARYA NAGARJUNA UNIVERSITY

CENTRE FOR DISTANCE EDUCATION

NAGARJUNA NAGAR,

GUNTUR

ANDHRA PRADESH

Lesson 1

BIVARIATE RANDOM VARIABLE

Objective:

After studying the lesson the students will have clear comprehension of the theory and practical utility about the concepts of bivariate random variable, joint probability function, marginal probability functions, joint probability distribution function, marginal probability distribution functions, joint probability density function, marginal probability density functions, conditional density functions, additive and multiplicative theorems of mathematical expectation.

Structure of the Lesson:

1.1 Introduction:

In the earlier study on probability and distributions, we have discussed univariate random variable and its probability distributions. Now, we extend our study to bivariate random variable.

The two jointly distributed univariate random variables on a same sample space, constitutes bivariate random variable. This type of variables often occur in practice. For example, the pressure and volume of gas, height and weight of students constitute bivariate random variable. This study is a fundamental basis for the concepts of curve fitting, correlation and regression.

So far we have taken only one measurement on a single item under observation. However, in many practical situations it is often possible and desirable to take more than one measurement on an item. Suppose we are interested in knowing the relation of area X and weight Y of leaves of a tree and collect data. We obtain data of the form (x_i, y_i) for $i = 1, 2, ..., n$. Such data is referred as bivariate data and the distribution as bivariate distribution. (X, Y) is called bivariate random variable taking values (x_i, y_i) for $i = 1, 2, ..., n$ on n items.

1.2 Bivariate Random Variable:

Let 'S' be the sample space associated with a given random experiment. A real valued function defined on 'S' and taking values in $\mathbb{R}\ (-\infty, \infty)$ is called one - dimensional random variable. If the functional values are ordered pairs of real numbers i.e. vectors in two-space the function is said to be two-dimensional random variable or bivariate random variable.

1.3 Joint Probability Function of X and Y:

Let X and Y be random variables on a sample space S with respective image sets $X(S) = \{x_1, x_2, ..., x_n\}$ and $Y(S) = \{y_1, y_2, ..., y_m\}$ we make the product set.

$$X(S) \times Y(S) = \{x_1, x_2, ..., x_n\} \times \{y_1, y_2, ..., y_m\}$$

into a probability space defining the probability of the ordered pair (x_i, y_j) to be $P(X = x_i, Y = y_j)$ simply $P(x_i, y_j)$. The function 'P' on $X(S) \times Y(S)$ defined by

$p_{ij} = P(X = x_i \cap Y = y_j) = P(x_i, y_j)$ is called the joint probability function of X and Y. It can be represented in the following table.

X \ Y	y_1	y_2	....	y_j	....	y_m	Total
x_1	p_{11}	p_{12}	....	p_{1j}	....	p_{1m}	$p_{1\cdot}$
x_2	p_{21}	p_{22}	....	p_{2j}	....	p_{2m}	$p_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	p_{i1}	p_{i2}	....	p_{ij}	....	p_{im}	$p_{i\cdot}$
. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .
x_n	p_{n1}	p_{n2}	....	p_{nj}	....	p_{nm}	$p_{n\cdot}$
Total	$p_{\cdot 1}$	$p_{\cdot 2}$	....	$p_{\cdot j}$	....	$p_{\cdot m}$	1

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} = 1 \cdot$$

1.4 Marginal Probability Functions of X and Y:

Suppose the joint distribution of two random variables X and Y is given, the probability distribution of X is as follows:

$$P(X=x_i) = P_X(x_i) = P(X=x_i \cap Y=y_1)+P(X=x_i \cap Y=y_2)$$

$$+...............+ P(X=x_i \cap Y=y_m)$$

$$=p_{i1}+p_{i2}+......+p_{im}$$

$= \sum_{j=1}^{m} p_{ij} = p_{i\cdot}$ and is known as marginal probability function of X. Also

$$\sum_{i=1}^{n} p_{i\cdot} = p_{1\cdot}+p_{2\cdot}+........+p_{n\cdot} = 1$$

Suppose the joint distribution of two random variables X and Y is given, then the probability distribution of Y is as follows:

$$P(Y=y_j)=P_Y(y_j)=P(X=x_1 \cap Y=y_j)+P(X=x_2 \cap Y=y_j)+.............+P(X=x_n \cap Y=y_j)$$

$$=p_{1j}+p_{2j}+..........+p_{nj}$$

$$= \sum_{i=1}^{n} p_{ij} = p_{\cdot j}$$

and also $\sum_{j=1}^{m} p_{\cdot j} = 1$.

Example 1: X and Y are two random variables having joint density function $f(x,y)=\frac{(x+2y)}{27}$, where X and Y can assume only the integers 0, 1, 2. Find the joint probability function of X and Y in the table and the marginal distributions of X and Y.

Solution : Joint density function of X and Y in the table as follows:

X \ Y	0	1	2	Total $P(x)$
0	0	$2/27$	$4/27$	$6/27$
1	$1/27$	$3/27$	$5/27$	$9/27$
2	$2/27$	$4/27$	$6/27$	$12/27$
Total $P(y)$	$3/27$	$9/27$	$15/27$	1

Marginal probability function of X

X	0	1	2
P(x)	6/27	9/27	12/27

Marginal probability function of Y

Y	0	1	2
P(y)	3/27	9/27	15/27

1.5 Conditional Probability Functions:

Conditional probability function of X given y = y_j is $P\left(X = x_i \middle| Y = y_j\right)$ defined as

$$P\left(X = x_i \middle| Y = y_j\right) = \frac{P\left(X = x_i \cap Y = y_j\right)}{P\left(Y = y_j\right)}$$

$$= \frac{P\left(x_i, y_j\right)}{P\left(y_j\right)} = \frac{p_{ij}}{p_{\cdot j}}$$

Where $P\left(x_i, y_j\right)$ is the joint probability function of X and Y and $P\left(y_j\right)$ is the marginal probability function of Y.

Conditional probability function of Y given $X = x_i$ is

$P\left(Y=y_j\middle|\,X=x_i\right)$ defined as $P\left(Y=y_j\middle|\,X=x_i\right)=\dfrac{P\left(X=x_i\cap Y=y_j\right)}{P\left(X=x_i\right)}$

$$=\frac{P\left(x_i,y_j\right)}{P\left(x_i\right)}=\frac{p_{ij}}{p_{i\cdot}}$$

Where $P\left(x_i\right)$ is the marginal probability function of X.

Note: (1) $\displaystyle\sum_{i=1}^{n}\frac{p_{ij}}{p_{\cdot j}}=\frac{p_{1j}+p_{2j}+\ldots\ldots\ldots\ldots+p_{nj}}{p_{\cdot j}}=\frac{p_{\cdot j}}{p_{\cdot j}}=1,$

similarly (2) $\displaystyle\sum_{j=1}^{m}\frac{p_{ij}}{p_{i\cdot}}=1\,.$

Example 2: For Example 1, find conditional distribution of Y for X = x.

Solution: Conditional probability for Y given $X=x_i$ is $P\left(Y=y_j\middle|X=x_i\right)$

$$P\left(\;Y=0\middle|\,X=0\right)=\frac{0}{6/27}=0$$

$$P\left(Y=0\middle|X=1\right)=\frac{1/27}{9/27}=1/9$$

$$P\left(Y=0\middle|X=2\right)=\frac{2/27}{12/27}=1/6$$

$$P\left(Y=2\middle|\,X=0\right)=2/3$$

$$P\left(Y=2\middle|\,X=1\right)=5/9$$

$$P\left(Y=2\middle|\,X=2\right)=1/2$$

$$P\left(Y=1\middle|\,X=0\right)=1/3$$

$$P\left(Y=1\middle|\,X=1\right)=1/3$$

$$P\left(Y=1\middle|\,X=2\right)=1/3$$

The above calculations are shown in the following table. Conditional probability for Y given $X = x_i$

Y X	0	1	2
0	0	$1/3$	$2/3$
1	$1/9$	$1/3$	$5/9$
2	$1/6$	$1/3$	$1/2$

1.6 Independence of two random variables X and Y:

Two random variables X and Y are said to be independent if

$$P\left(X = x_i,\ Y = y_j\right) = P\left(X = x_i\right) \cdot P\left(Y = y_j\right),$$

otherwise they are said to be dependent,

where $P\left(X = x_i,\ Y = y_j\right)$ is joint probability function of X and Y.

$P\left(X = x_i\right)$ is the marginal probability function of X.

$P\left(Y = y_j\right)$ is the marginal probability function of Y.

1.7 Joint Probability Distribution Function:

Let (X, Y) be a two dimensional (bivariate) random variable then their joint distribution function is denoted by $F_{XY}(x, y)$ and it represents the probability that simultaneously the observation (x, y) will have the property $(X \leq x,\ Y \leq y)$.

i.e., $F_{XY}(x, y) = P(-\infty < X \leq x,\ -\infty < Y \leq y)$

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

$$= \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y)\, dxdy \ ; \text{ where } f(x, y) \geq 0$$

and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy=1$ (for contionuous random variable)

$\sum_{x}\sum_{y} f(xy)=1$ (for discrete random variable)

Properties:

1(a) For the real numbers a_1, b_1, a_2 and b_2

$$P(a_1<X\le b_1,\ a_2<Y\le b_2)=F_{XY}(b_1,b_2)+F_{XY}(a_1,a_2)-F_{XY}(a_1,b_2)-F_{XY}(b_1,a_2)$$

(b) Let $a_1<a_2,\ b_1<b_2$ we have $F(a_1,b_2)\ge F(a_1,a_2)$

which shows that F(x,y) is monotonic non-decreasing function.

2. $F(-\infty,y)=O=F(x,\infty);\ F(-\infty,\infty)=1.$

3. If the density function f (x,y) is continuous at (x, y) then

$$\frac{\partial^2 F(x,y)}{\partial x\partial y}=f(x,y)\cdot$$

1.8 Marginal Probability Distribution Functions:

Marginal distribution function of X is

$$F_X(x)=P(X\le x)=P(X\le x,\ Y<\infty)=\underset{y\to\infty}{\mathrm{Lt}}\ F_{XY}(xy)=F_{XY}(x,\infty).$$

Marginal distribution of Y is

$$F_Y(y)=P(Y\le y)=P(X<\infty,\ Y\le y)=\underset{x\to\infty}{\mathrm{Lt}}\ P_{XY}(x,y)=F_{XY}(\infty,y).$$

In case of jointly discrete random variables, the marginal distribution functions of X and Y are given as

$$F_x(x)=\sum_{y}P(X=x,\ Y=y)$$

$$F_y(y)=\sum_{x}P(X=x,\ Y=y)$$

In case of jointly continuous random variables, the marginal distribution functions are given as

$$F_X(x) = \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f_{XY}(x,y)\, dy \right] dx$$

$$F_Y(y) = \int_{-\infty}^{y} \left[\int_{-\infty}^{\infty} f_{XY}(x,y)\, dx \right] dy.$$

1.9 Joint Probability density Function of X and Y:

The probability that point (x, y) will lie in the infinitesimal rectangular region of area dxdy is given by

$$P\left(x - \frac{dx}{2} \le X \le x + \frac{dx}{2},\ y - \frac{dy}{2} \le Y \le y + \frac{dy}{2} \right) = d F_{XY}(x,y)$$

and is denoted by $f_{XY}(xy)dxdy$ where function $f_{XY}(x,y)$ is called the joint density function of X and Y.

1.10 Marginal Probability density Functions of X and Y:

Marginal density function of X, $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(xy)dy$

(for continuous random variable)

$= \sum_{Y} P_{XY}(xy)$ (for discrete random variable)

and marginal density function of Y

$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx$ (for continuous random variable)

$= \sum_{X} P_{XY}(x,y)$ (for discrete random variable)

Also obtained in the following way.

$$f_X(x) = \frac{dF_X(x)}{dx} = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$$

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \int_{-\infty}^{\infty} f_{XY}(x,y)dx$$

1.11 Conditional distribution functions:

Conditional distribution function $F_{Y|X}(y|\, x)$ denotes the distribution of Y when X has already assumed the particular value x. Hence

$$F_{Y|X}(y|\, x) = P[Y \le y|\, X = x].$$

The joint distribution function $F_{XY}(x,y)$ may be expressed interms of the conditional distribution function as follows:

$$F_{XY}(x,y) = \int_{-\infty}^{x} F_{Y|X}(y\,|\,x)dF_X(x)$$

$$F_{XY}(x,y) = \int_{-\infty}^{y} F_{X|Y}(x\,|\,y)dF_Y(y)$$

1.12 Conditional density functions:

Random variables X and Y are jointly distributed. Conditional density function of Y given X is

$$f_{Y|X}(y\,|\,x) = \frac{\partial}{\partial y}F_{Y|X}(y\,|\,x) = \frac{f_{XY}(x,y)}{f_X(x)};\ f_X(x) > 0.$$

Conditional density function of X given Y is

$$f_{X|Y}(x\,|\,y) = \frac{\partial}{\partial x}F_{X|Y}(x\,|\,y) = \frac{f_{XY}(x,y)}{f_Y(y)};\ f_Y(y) > 0\ ,$$

where $f_{XY}(x,y)$ is the joint density function of X and Y,

$f_X(x)$ is the marginal density function of X,

and $f_Y(y)$ is the marginal density function of Y.

1.13 Stochastic Independence:

Two random variables X and Y with joint p.d.f. $f_{XY}(x,y)$ and marginal p.d.f's $f_X(x)$ and $f_Y(y)$ respectively are said to be stochastically independent if and only if

$$f_{XY}(x,y)=f_X(x)f_Y(y).$$

Example 3: If $f(x,y)=8xy,\ 0<x<y<1$

= 0 (otherwise)

Find marginal density functions of X and Y and check the independence of the random variables.

Solution: Marginal density function of $X=f_X(x)=\int_{-\infty}^{\infty} f_{XY}(x,y)dy$

$$=\int_0^1 8xy\,dy=8x\left[\frac{y^2}{2}\right]_0^1$$

$$=8x\cdot\left[\frac{1}{2}-0\right]=4x\ .$$

Marginal density function of $Y=f_Y(y)=\int_{-\infty}^{\infty} f_{XY}(x,y)dx$

$$=\int_0^y 8xydx$$

$$=8y\cdot\left[\frac{y^2}{2}\right]_0^y$$

$$=8y\left[\frac{y^2}{2}\right]=4y^3$$

According to the definition of independence

$$f_{XY}(x,y) = f_X(x) \cdot f_Y(y),$$

but $8xy \neq 4x \cdot 4y^3 = 16xy^3$

$\therefore$ X and Y are not independent.

1.14 Covariance:

If X and Y are two random variables then covariance between them is defined as

$$\text{Cov}(X,Y) = E\{[X - E(X)][Y - E(Y)]\} = E(XY) - E(X)E(Y)$$

If X and Y are independent then E(XY) = E(X) E(Y) and hence Cov (X, Y) = 0.

1.15 Addition Theorem of Mathematical Expectation:

If X and Y are random variables then $E(X+Y) = E(X) + E(Y)$ provided all the expectations exist.

Proof:

Let X and Y be continuous r.v.'s with joint p.d.f. $f_{XY}(x,y)$ and marginal p.d.f.'s $f_X(x)$ and $f_Y(y)$ respectively. Then by definition

$$E(X) = \int_{-\infty}^{\infty} x \quad f_X(x)dx \cdots\cdots\cdots\cdots\cdots\cdots(1)$$

$$E(Y) = \int_{-\infty}^{\infty} y \, f_Y(y)dy \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(2)$$

$$E(X+Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y) f_{XY}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \, f_{XY}(x,y) \, dxdy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \, f_{XY}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{XY}(x,y)dy \right] dx + \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} f_{XY}(x,y)dx \right] dy$$

$$= \int_{-\infty}^{\infty} x f_X(x)dx + \int_{-\infty}^{\infty} y\, f_Y(y)dy = E(X) + E(Y) \text{ (from (1) and (2))}$$

The result can be extended to n variables as

$$E(X_1 + X_2 + ... + X_n) = E(X_1) + E(X_2) + ... + E(X_n).$$

1.16 Multiplication Theorem of Mathematical Expectation:

If X and Y are independent random variables, then

$$E(XY) = E(X) \cdot E(Y)$$

Proof: $E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, f_{XY}(x,y)\, dxdy$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f_X(x) \cdot f_Y(y)\, dxdy \text{ (Since X and Y are independent)}$$

$$= \int_{-\infty}^{\infty} x f_X(x)dx \cdot \int_{-\infty}^{\infty} y\, f_Y(y)dy$$

$$= E(X) \cdot E(Y).$$

The result can be extended to 'n' variables as

$$E(X_1\ X_2 ....X_n) = E(X_1) \cdot E(X_2)..................E(X_n).$$

Example 4: Two random variables X and Y have the following p.d.f.

$f(x,y) = 2 - x - y;\ 0 \le x \le 1,\ 0 \le y \le 1$

= 0 (otherwise)

Find (i) Marginal p.d.f's of X and Y

(ii) Conditional density functions

(iii) Var(X) and Var(Y)

(iv) Covariance between X and Y

Solution: (i) Marginal p.d.f's:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^1 (2-x-y)\,dy = \frac{3}{2} - x\ ;\ 0 < x < 1$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_0^1 (2-x-y)\,dx = \frac{3}{2} - y\ ;\ 0 < y < 1$$

(ii) Conditional density functions:

$$f_{X|Y}(x\,|\,y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{2-x-y}{\left(3/2 - y\right)}\ ;\ 0 < (x,y) < 1$$

$$f_{Y|X}(y\,|\,x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{2-x-y}{\left(3/2 - x\right)} ; 0 < (x,y) < 1$$

(iii) $$E(X) = \int_0^1 x\ f_X(x)dx = \int_0^1 x\left(\frac{3}{2} - x\right)dx = \frac{5}{12}$$

$$E(Y) = \int_0^1 y\ f_Y(y)dy = \int_0^1 y\left(\frac{3}{2} - y\right)dy = \frac{5}{12}$$

$$E\left(X^2\right) = \int_0^1 x^2 f_X(x)\,dx = \int_0^1 x^2\left(\frac{3}{2} - x\right)dx = \frac{1}{4}$$

$$E\left(Y^2\right) = \int_0^1 y^2\ f_Y(y)\,dy = \int_0^1 y^2\left(\frac{3}{2} - y\right)dy = \frac{1}{4}$$

$$V(X) = E\left(X^2\right) - \left[E(X)\right]^2 = \frac{1}{4} - \left(\frac{5}{12}\right)^2 = \frac{11}{144}$$

Similarly $V(Y) = E\left(Y^2\right) - \left[E(Y)\right]^2 = \frac{11}{144}$

(iv) $$E(X,Y) = \int_0^1 \int_0^1 xy\,(2-x-y)\,dxdy$$

$$= \int_0^1 \left[2\ \frac{x^2y}{2} - \frac{x^3y}{3} - \frac{x^2y^2}{2}\right]_0^1 dy$$

$$= \int_0^1 \left(\frac{2}{3}y - \frac{1}{2}y^2\right)dy$$

$$= \frac{1}{6}\cdot$$

$$\text{Cov}\,(XY) = E(XY) - E(X)E(Y)$$

$$= \frac{1}{6} - \frac{5}{12}\cdot\frac{5}{12} = -\frac{1}{144}\cdot$$

1.17 Exercises:

(1) Joint probability mass function of (x,y) is given in the following table. Find (i) Marginal distribution of 'X' , (ii) E(XY) and E(X), E(Y), (iii) V(X); V(Y), (iv) Cov (X,Y), (v) r(X,Y).

Y X	0	1	2
1	2/24	3/24	0
3	0	4/24	0
5	1/24	2/24	6/24
7	3/24	2/24	1/24

(2) If $f(x,y) = K(4-x-y);\ 0 \le x \le 2,\ 0 \le y \le 2$

$= 0$ (otherwise)

(i) Find the constant K. (ii) Marginal density functions of X and Y.

(3) The joint p.d.f. of bivariate r.v is given below

$$f(x,y) = \frac{1}{20}(2x+3)\,e^{-y/2};\ 0 \le x \le 2;\ y \ge 0$$

$= 0$ (otherwise)

Find (i) Mean values of X and Y (ii) V(X), V(Y)

(iii) Cov (X, Y) (iv) r(X, Y)

(4) If $f(x,y) = \frac{2}{5}(x+2)\,e^{-y}\ ;\ 0 < x < 1,\ y > 0$

= 0 (otherwise)

Derive the marginal p.d.f.'s and comment.

(5) If $f(x,y) = \frac{1}{8}(6-x-y);\ 0 \le x \le 2,\ 2 \le y \le 4$

$= 0$ (otherwise)

Find the r(X, Y).

1.18 Summary:

Concepts of bivariate random variable, joint probability function of X and Y, marginal probability functions of X and Y, conditional probability functions, independence of two random variables, joint probability distribution function, marginal probability distribution functions, conditional distribution function, conditional density functions are discussed. Some problems are solved to illustrate some of the concepts and some problems are given as exercises to the students to solve on their own.

1.19 Technical Terms:

Bivariate random variable, Joint Probability function, conditional probability functions, Independence of random variables.

Lesson Writer

A. Mohan Rao

Lesson 2

CURVE FITTING

Objective:

After studying the lesson the student will be able to understand about scatter diagram and method of least squares for fitting various curves such as straight line, quadratic, power and exponential curves for a given data.

Structure of the Lesson:

2.1 Introduction:

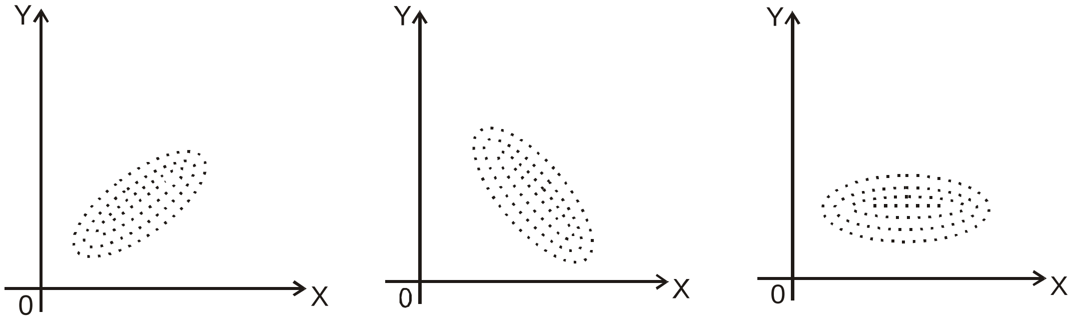
An important application of statistics is to predict a future value of Y dependent on a set of related independent variables $X_1, X_2, ..., X_n$. Thus we set up a model which relates to the dependent variable Y to the independent variable values $X_1, X_2, ..., X_n$. We shall fit the model by method of least squars.

2.2 Scatter diagram:

It is the simplest way of the diagrammatic representation of bivariate data. Thus for the bivariate distribution $(x_i, y_i); i = 1, 2, ..., n$ if the values of the variables X and Y be plotted along the X-axis and Y-axis respectively in the xy plane, the diagram of dots so obtained is known as scatter diagram, since the term scatter refers to the dispersion of dots on the graph.

From the scatter diagram, we can form a fairly good, though vague, idea whether the variables are correlated or not. e.g., if the points are very dense i.e. very close to each other, we should expect a fairly good amount of correlation between the variables and if the points are widely scattered, a poor correlation is expected. However, this method is not suitable if the number of observations is fairly large.

Following are the figures of scattered data



Positive Correlation Negative Correlation No Correlation

2.3 Principle of Least Squares:

Minimising the sum of squares of the deviations of the actual values of 'y' from its estimated values $(\hat{y})$ i.e. $\sum_{i=1}^{n}(y_i - \hat{y})^2$ is minimum.

The difference between the observed values and expected values is known as residual or error and the task is to minimise these results. Since these differences may be positive or negative, it is more convenient to make the sum of squares of these residuals minimum. This is known as the method of least squares. If o_i be the observed values and e_i be the expected values then $\sum_{i=1}^{n}(o_i - e_i)^2$ is minimum.

2.4 Fitting of Straight Line:

Let us consider the fitting of a straight line

$$Y = a + bX \rightarrow (1)$$

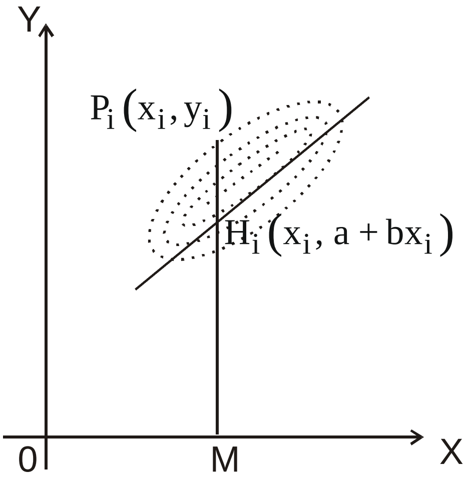
to a set of n points $(x_i, y_i); i=1,2,..........,n$. Equation (1) represents a family of straight lines for different values of the arbitary constants 'a' and 'b'. The problem is to determine 'a' and 'b' so that the line (1) is the line of "best fit".

Let $P_i(x_i, y_i)$ be any general point in the scatter diagram. Draw P_iM perpendicular to the x-axis meeting the straight line equation (1) in H_i. Abscissa of H_i is x_i and since H_i lies on (1) its ordinate is $a + bx_i$. Hence the co-ordinates of H_i are $(x_i, a + bx_i)$.

$P_iH_i = P_iM - MH_i$

$= y_i - a - bx_i$, is called error of estimate or the residual for y.

According to the principle of least squares we have to determine 'a' and 'b' so that $E = \sum_{i=1}^{n} P_iH_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$ is minimum.



From the principle of maximam and minimam the partial derivatives of 'E' w.r.t 'a' and 'b' should vanish separetely i.e. differentiating 'E' w.r.t. 'a' partially and equating to '0'.

$$\frac{\partial E}{\partial a} = 0 = -2\sum_{i=1}^{n} (y_i - a - bx_i)$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \rightarrow (2)$$

Differentiating 'E' w.r.t. 'b' partially and equating to '0'

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

$$\Rightarrow \sum_{i=1}^{n} x_iy_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \rightarrow (3)$$

Equations (2) and (3) are called normal equations for estimating 'a' and 'b'.

All the quantities required in equation (2) and (3) to estimate 'a' and 'b' can be obtained from the given set of points $\left(x_i, y_i\right);\ i=1,2,........,n$. With the values so obtained equation (1) is the straight line of the best fit to the given set of points. i.e. $Y=\hat{a}+\hat{b}X$; Where $\hat{a}$ and $\hat{b}$ are the estimated values of 'a' and 'b'.

2.5 Fitting of Quadratic Curve or Second degree parabola:

Let $Y=a+bX+cX^2 \rightarrow (1)\ \ (c \neq 0)$ be the second degree parabola of best fit to the set of 'n' points $\left(x_i, y_i\right);\ i=1,2,.............,n$. Using the method of least squares, we have to determine the arbitrary constants a, b and c, so that

$$E=\sum_{i=1}^{n}\left(y_i-a-bx_i-cx_i^{\,2}\right)^2 \quad \text{is minimum.}$$

Equating to zero, the partial derivatives of 'E' w.r.t. a, b and c separately, we get

$$\frac{\partial E}{\partial a}=0=-2\sum_{i=1}^{n}\left(y_i-a-bx_i-cx_i^{\,2}\right)$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^{\,2} \rightarrow (2)$$

$$\frac{\partial E}{\partial b}=0=-2\sum_{i=1}^{n}\left(y_i-a-bx_i-cx_i^{\,2}\right)x_i$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} + c\sum_{i=1}^{n} x_i^{\,3} \rightarrow (3)$$

$$\frac{\partial E}{\partial c}=0=-2\sum_{i=1}^{n}\left(y_i-a-bx_i-cx_i^{\,2}\right)x_i^2$$

$$\Rightarrow \sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \rightarrow (4)$$

Equations (2), (3) and (4) are known as normal equations for estimating 'a', 'b' and 'c'. For the given set of points $\left(x_i,\ y_i\right);\ i=1,2,..........,n$, equations (2), (3) and (4) can be solved for a, b and c and estimated values are $\hat{a},\hat{b}$ and $\hat{c}$. With these values equation (1) is the required parabola

i.e. $Y=\hat{a}+\hat{b}X+\hat{c}X^2$.

2.6 Fitting of Power Curve:

Let $Y = aX^b \rightarrow (1)$ be the power curve to be fitted to a set of n points $(x_i, y_i); i=1,2,........., n$.

Taking logarithms on both sides of (1) we get

$$\log Y = \log a + b \log X$$

$$\Rightarrow U = A + bV \rightarrow (2)$$

where $U = \log Y; A = \log a ; V = \log X$

so that equation (2) is the linear equation in U and V.

The normal equations are

$$\sum_{i=1}^{n} U_i = nA + b\sum_{i=1}^{n} V_i \rightarrow (3)$$

$$\sum_{i=1}^{n} U_i V_i = A\sum_{i=1}^{n} V_i + b\sum_{i=1}^{n} V_i^2 \rightarrow (4)$$

Solving (3) and (4) for A and b; we get $\hat{a}$ = antilog $\hat{A}, \hat{b}$.

With the estimated values $\hat{a}$ and $\hat{b}$, equation (1) is the curve of the best fit to the given set of 'n' points $(x_i, y_i); i=1,2,........., n$.

i.e. $Y = \hat{a}\, X^{\hat{b}}$.

2.7 Fitting of Exponential Curves:

(i) Let $Y = ab^X \rightarrow (1)$ be the exponential curve to be fitted to a set of 'n' points, $(x_i, y_i); i=1,2,........., n$. Taking logarithms on both sides, we get

$$\log Y = \log a + X \log b$$

$$\Rightarrow U = A + BX \rightarrow (2)$$

where A = log a; B = log b; U = log Y.

Equation (2) is the linear equation in X and U.

The normal equations for estimating A and B are

$$\sum_{i=1}^{n} U_i = nA + B\sum_{i=1}^{n} X_i \rightarrow (3)$$

$$\sum_{i=1}^{n} U_i X_i = A\sum_{i=1}^{n} X_i + B\sum_{i=1}^{n} {X_i}^2 \rightarrow (4)$$

Solving equations (3) and (4) for A and B, we get

$$\hat{a} = \text{anti}\log \hat{A}$$

$$\hat{b} = \text{anti}\log \hat{B}$$

With these estimated values of a and b equation (1)

$Y = \hat{a}\hat{b}^X$ is the curve of the best fit.

(ii) Exponential Curve of the type $Y = ae^{bX}$.

Let $Y = ae^{bX} \rightarrow (1)$ be the exponential curve to be fitted to a set of n points $(x_i, y_i);\ i = 1, 2, \ldots\ldots\ldots, n$.

Taking logarithms on both sides, we get

$$\log Y = \log a + bX. \log e$$

$$\Rightarrow U = A + BX \rightarrow (2)$$

where A = log a; B = b log e; U = log Y.

Equation (2) is a linear equation in X and U.

The normal equations for estimating A and B are

$$\sum_{i=1}^{n} U_i = nA + B\sum_{i=1}^{n} X_i \rightarrow (3)$$

$$\sum_{i=1}^{n} U_i X_i = A\sum_{i=1}^{n} X_i + B\sum_{i=1}^{n} {X_i}^2 \rightarrow (4)$$

Solving (3) and (4) for A and B, we get

$$\hat{a} = \text{Anti}\log \hat{A}$$

$$\hat{b} = \frac{\hat{B}}{\log e}$$

With these values of $\hat{a}$ and $\hat{b}$ equation (1) is the best fit to the given set of 'n' points.

i.e. $Y = \hat{a}\, e^{\hat{b}X}$.

2.8 Problems:

Problem 1: Fit a straight line to the following data with 'x' as the independent variable.

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

Solution: Let the straight to be fitted to the given data is

$$Y = a + bX \rightarrow (1)$$

Normal equations are

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

X	Y	XY	X^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
$\sum X = 10$	$\sum Y = 16.9$	$\sum XY = 47.1$	$\sum X^2 = 30$

$$n = 5;\ \sum X = 10;\ \sum Y = 16.9;\ \sum XY = 47.1;\ \sum X^2 = 30$$

Substituting these values in normal equations, we get

$$16.9 = 5a + 10b \rightarrow (2)$$

$$47.1 = 10a + 30b \rightarrow (3)$$

Solving these equations, we get a = 0.72, b = 1.33.

Thus the required straight line is Y = 0.72 + 1.33X.

Problem 2: Fit a second degree parabola to the following data

X	1	2	3	4	5
Y	6	17	34	57	86

Solution: Let $Y = a + bX + cX^2 \rightarrow (1)$ be the second degree parabola to be fitted to the given data and normal equations for estimating a, b and c are

$$\sum Y = na + b\sum X + c\sum X^2$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$

$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

X	Y	XY	X^2	X^3	X^4	X^2Y
1	6	6	1	1	1	6
2	17	34	4	8	16	68
3	34	102	9	27	81	306
4	57	228	16	64	256	912
5	86	430	25	125	625	2150
15 $\sum X =$	200 $\sum Y =$	800 $\sum XY =$	55 $\sum X^2 =$	225 $\sum X^3 =$	979 $\sum X^4 =$	3442 $\sum X^2Y =$

and n = 5; substituting these values in the normal equations, we get

$$5a + 15b + 55c = 200 \rightarrow (2)$$

$$15a + 55b + 225c = 800 \rightarrow (3)$$

$$55a + 225b + 979c = 3442 \rightarrow (4)$$

Solving these equations we get

$$a = 1, \; b = 2, \; c = 3$$

Thus the required equation is $Y = 1 + 2X + 3X^2$.

Change of Origin: Let us suppose that the values of X are given to be equidistant at an interval of h i.e. 'X' takes the values (say) $a\,,\ a+h, a+2h,$.................. If 'n' is odd i.e. $n=2m+1$ (say) we take

$$U=\frac{X-(\text{Middle term})}{\text{Interval}}=\frac{X-(a+mh)}{h}$$

Now 'U' takes the values $-m,\ -(m-1),........,-1,0,1,...............,(m-1),m,$ so that $\sum U=\sum U^3=0$.

If 'n' is even i.e. n = 2m (say), then there are two middle terms, viz, m^{th} and $(m+1)^{th}$ terms which are $\{a+(m-1)h\}$ and $(a+mh)$. In this case we take

$$U=\frac{X-(\text{Mean of two middle terms})}{\frac{1}{2}(\text{Interval})}=\frac{X-\frac{\{[a+(m-1)h]+[a+mh]\}}{2}}{\frac{1}{2}(h)}$$

Now 'U' takes the values

$$-(2m-1),\ -(2m-3),..............,-3,-1,1,3,.............,(2m-3),(2m-1).$$

Again we see that $\sum U=\sum U^3=0$.

Problem 3: Fit a straight line of the form $Y=aX+b$ to the following data.

X	0	5	10	15	20	25	30
Y	10	14	19	25	31	36	39

Solution: In this we use the change of origin and scale as follows $U=\frac{X-a}{h}$; $V=Y-b$, then normal equations are

$$\sum V=nb+a\sum U \rightarrow (1)$$

$$\sum UV=b\sum U+a\sum U^2 \rightarrow (2)$$

X	Y	$U=\frac{X-15}{5}$	$V=Y-19$	UV	U^2
0	10	-3	-9	27	9
5	14	-2	-5	10	4
10	19	-1	0	0	1
15	25	0	6	0	0
20	31	1	12	12	1
25	36	2	17	34	4
30	39	3	20	60	9
		$\sum U=0$	$\sum V=41$	$\sum UV=143$	$\sum U^2=28$

n = 7;

From (1) $7b+0=41 \Rightarrow b=\frac{41}{7}=5.87$

From (2) $28\,a=143 \Rightarrow a=\frac{143}{28}=5.11$

$$V=5.11U+5.87;\ \text{but}\ V=Y-19;\ U=\frac{X-15}{5}$$

$$Y-19=5.11\frac{(X-15)}{5}+5.87$$

$\Rightarrow Y=1.02X+9.54$ is the best fitted straight line to the given data.

Problem 4: Fit a second degree parabola $Y=a+bX+cX^2$ to the following data relating to profit of a certain country.

Year (t) :	1960	1962	1964	1966	1968
Profit (Y):	125	140	165	195	230

Solution: Second degree parabola $Y=a+bX+cX^2 \rightarrow (1)$ and normal equations are

$$\sum Y=na+b\sum X+c\sum X^2$$

$$\sum XY=a\sum X+b\sum X^2+c\sum X^3$$

$$\sum X^2Y=a\sum X^2+b\sum X^3+c\sum X^4.$$

Year (t)	Profit (y)	$X = \frac{t-1964}{2}$	XY	X^2Y	X^2	X^3	X^4
1960	125	-2	-250	500	4	-8	16
1962	140	-1	-140	140	1	-1	1
1964	165	0	0	0	0	0	0
1966	195	1	195	195	1	1	1
1968	230	2	460	920	4	8	16
	$\sum Y = 855$	$\sum X = 0$	$\sum XY =$ 265	$\sum X^2Y =$ 1755	$\sum X^2 =$ 10	$\sum X^3 =$ 0	$\sum X^4 =$ 34

Here n = 5

Substituting above values in the normal equations

$$855 = 5a + 10c \rightarrow (2)$$

$$265 = 10b \rightarrow (3)$$

$$1755 = 10a + 34c \rightarrow (4)$$

Solving (2), (3) and (4) we get a = 164.6; b = 26.5; c=3.2

The best fitted second degree curve to the given data is

$$Y = 164.6 + 26.5X + 3.2X^2$$

Problem 5: Fit a power curve to the following data.

X	1	2	3	4	5
Y	3	12	27	48	75

Solution: Let $Y = aX^b$ be the power curve .... (1)

$$\log Y = \log a + b\log X$$

$\Rightarrow$ $U = A + bV \rightarrow$ is the linear equation. Normal equations are

$$\sum U = nA + b\sum V \; ; \; \sum UV = A\sum U + b\sum V^2$$

X	Y	U = log Y	V = log X	V^2	UV
1	3	0.4771	0	0	0
2	12	1.0792	0.3010	0.0906	0.3248
3	27	1.4314	0.4771	0.2276	0.6829
4	48	1.6812	0.6021	0.3625	1.0122
5	75	1.8751	0.6990	0.4886	1.3108
		6.5440 $\sum U =$	2.0792 $\sum V =$	1.1693 $\sum V^2 =$	3.3307 $\sum UV =$

$$6.5440 = 5A + 2.0792b \rightarrow (2)$$

$$3.3307 = 2.0792A + 1.1693b \rightarrow (3)$$

$$(2) \times 2.0792 \Rightarrow 13.6063 = 10.396A + 4.3231b$$

$$(3) \times 5 \qquad \Rightarrow 16.6535 = 10.396A + 5.8468b$$

$$3.0472 = \qquad 1.5237b$$

$$b = \frac{3.0472}{1.5237} = 1.99989$$

Substituting b = 1.99989 in (2)

$$A = \frac{6.544 - 2.0792 \times 1.99989}{5} = 0.477165$$

$$a = \text{Anti}\log\ A = \text{Anti}\log\ 0.477165 = 3.0003$$

$$\therefore a = 3.0003 \cong 3$$

$$b = 1.99989 \cong 2$$

$\therefore$ The best fitted power curve $Y = 3X^2$

Problem 6: Fit an exponential curve of the type $Y = ab^X$ to the following data. Estimate Y when X = 3.5

X	2	3	4	5	6
Y	144	172.8	207.4	248.8	298.6

Solution: Let $Y = ab^X \rightarrow (1)$

Taking log and using the exponential method, we get

$U = A + BX \rightarrow (2)$ and the normal equations are

$$\sum U = nA + B\sum X$$

$$\sum UX = A\sum X + B\sum X^2$$

X	Y	U = log Y	UX	X^2
2	144	2.1584	4.3168	4
3	172.8	2.2375	6.7125	9
4	207.4	2.3168	9.2672	16
5	248.8	2.3959	11.9795	25
6	298.6	2.4751	14.8506	36
20		11.5837	47.1266	90

$\sum X =$ 20; $\sum U =$ 11.5837, $\sum UX =$ 47.1266, $\sum X^2 = 90$

$$5A + 20B = 11.5837 \quad \rightarrow (3)$$

$$20A + 90B = 47.1266 \rightarrow (4)$$

$(3) \times 4 \Rightarrow 20A + 80B = 46.3348$

10B = 0.7918

$$B = \frac{0.7918}{10} = 0.07918$$

Substitute B = 0.07918 in (3) we get A = 1.9999 $\cong$ 2

$a = \text{Anti}\log A = \text{anti}\log 2 = 100$

$b = \text{Anti}\log B = \text{Anti}\log (0.0792) = 1.2$

$\therefore$ The best fitted exponential curve to the given data is $Y = 100(1.2)^X$.

When X = 3.5

$$\log Y = 2.0 + 3.5 \times 0.0792$$

$$= 2.2772$$

$$Y = \text{Anti}\log(2.2772) = 189.3$$

$\therefore$ When X = 3.5, Y = 189.3

Problem 7: For the data given below find equation to the best fitting exponential curve of the form $Y = ae^{bX}$

X	1	2	3	4	5	6
Y	1.6	4.5	13.8	40.2	125.0	300

Solution: Let $Y = ae^{bX} \rightarrow (1)$

Taking log and using exponential method, we get

$V = A + BX \rightarrow (2)$ and the normal equations are

$$\sum V = nA + B\sum X$$

$$\sum VX = A\sum X + B\sum X^2$$

X	Y	V = log Y	VX	X^2
1	1.6	0.2041	0.2041	1
2	4.5	0.6532	1.3064	4
3	13.8	1.1399	3.1497	9
4	40.2	1.6042	6.4168	16
5	125.0	2.0969	10.4845	25
6	300.0	2.4771	14.8626	36
21		8.1754	36.6941	91

$\sum X =$ 21, $\sum V =$ 8.1754, $\sum VX = 36.6941$, $\sum X^2 =$ 91.

$$6A + 21B = 8.1754 \rightarrow (3)$$

$$21A + 91B = 36.6947 \rightarrow (4)$$

$$(3)\big/_2 \times 7 \Rightarrow \underline{21A + 73.5B = 28.6139}$$

$$17.5B = 8.0802$$

$$\Rightarrow B = 0.4617$$

Substitute B = 0.4617 in (3) $A = \dfrac{8.1754 - 2 \times 0.4617}{6}$

$$= -0.2534$$

$$b = \frac{B}{\log_{10} e} = \frac{0.4617}{0.4343} = 1.0752$$

a = Antilog (-0.2534) = 0.558

$\therefore$ The best fitted exponential curve to the given data is Y = (0.558) $e^{(1.0752)X}$

2.9 Model Questions and Exercises:

1. Derive the normal equations to fit a curve of the type $Y = ae^{bX}$, using least square principle.

2. Fit a second degree parabola to the following data.

X	0	1	2	3	4
Y	1	1.8	1.3	2.5	6.3

3. Derive the normal eqations for fitting the Curve $Y = aX^b$ by the principle of least squares.

4. Explain the principle of least squares. How do you fit a straight line to the given data.

5. Estimate the production for the year 2000 by fitting a straight line to the following data.

Year	1991	1992	1993	1994	1995	1996
Production	8	12	14	20	25	30

2.10 Summary:

In this lesson we have discussed about scatter diagram and method of least squares for fitting various curves such as straight line, quadratic, power and exponential curves for a given data. A good number of problems are solved for illustrating the fitting of different curves and some exercises are left over to the student.

2.11 Technical Terms:

Residual, Least squares, Normal equations, Power curves, exponential curves.

Lesson Writer

A. Mohan Rao

Lesson 3

CORRELATION

Objective:

After studying the lesson the students will be conversant with the concepts and applications of correlation, partial and multiple correlation coefficients, Spearman's rank correlation coefficient.

Structure of the Lesson:

3.1 Introduction:

In practice, we come across problems involving two or more variables. If we carfully study the data of rain fall and production of paddy, number of accidents and number of cars in a city, we may find that there is some relationship between two variables.

Two variables are said to be correlated if change in the value of one appears to be releted or linked with the change in the other one.

Example:

(i) The heights and weights of a group of persons

(ii) Income and expenditure

(iii) The volume and pressure of a perfect gas

3.2 Product moment correlation coefficient and its properties:

Product moment correlation coefficient: Karl Pearson's coefficient of correlation is also called product moment correlation coefficient.

As a measure of intensity or degree of linear relationship between two variables, Karl Pearson developed a formula called correlation coefficient.

Correlation coefficient between two random variables X and Y, usually denoted by r(X,Y) or simply r_{XY} is a numerical measure of linear relationship between them and is defined as $r(X,Y)=\frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$.

If $(x_i, y_i);\ i=1,2,.........,n$ is the bivariate distribution, then

$$\text{Cov}(X,Y)=E\{(X-E(X))(Y-E(Y))\}=\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})$$

$$=\frac{1}{n}\sum x_i y_i-\bar{x}\,\bar{y}$$

$$\sigma_x^2=E\{X-E(X)\}^2=\frac{1}{n}\sum(x_i-\bar{x})^2=\frac{1}{n}\sum x_i^2-(\bar{x})^2$$

$$\sigma_y^2=E\{Y-E(Y)\}^2=\frac{1}{n}\sum(y_i-\bar{y})^2=\frac{1}{n}\sum y_i^2-(\bar{y})^2$$

Assumptions in Karl Perarson's Coefficient of Correlation:

(i) The variables are random and are linearly related

(ii) The variables are effected by a number of independent causes having some inter related effects.

Properties:

(i) Limits for correlation coefficient

$$-1\le r(X,Y)\le 1$$

(ii) Correlation coefficient is independent of change of origin and scale.

(iii) Two independent variables are uncorrelated.

(i) Limits for correlation coefficient are $-1\le r(X,Y)\le 1$

Proof:

$$r(X,Y)=\frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\left[\frac{1}{n}\sum(x_i-\bar{x})^2\cdot\frac{1}{n}\sum(y_i-\bar{y})^2\right]^{1/2}}$$

$$\therefore r^2(X,Y)=\frac{\left(\sum a_i b_i\right)^2}{\left(\sum a_i^2\right)\left(\sum b_i^2\right)} \quad \text{where} \quad \begin{matrix} a_i = x_i - \bar{x} \\ b_i = y_i - \bar{y} \end{matrix} \rightarrow (1)$$

We have the Schwartz inequality which states that if $a_i, b_i \quad (i=1,2,.........,n)$ are real quantities, then

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \le \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

The sign of equality holding if and only if

$$\frac{a_1}{b_1}=\frac{a_2}{b_2}=..........=\frac{a_k}{b_k}.$$

Using Schwartz inequality in (1), we get

$r^2(X,Y)\le 1$, i.e. $|r(X,Y)|\le 1$

$\Rightarrow -1\le r(X,Y)\le 1$

(ii) Correlation coefficient is independent of change of origin and scale.

Proof:

Let$U=\frac{X-a}{h}$; $V=\frac{Y-b}{k}$; so that $X=a+hU$ and

$Y=b+kV$ where, a, b, h, k are constants, $h>0$, $k>0$.

We shall prove that $r(X,Y)=r(U,V)$

Since $X=a+hU$ and $Y=b+kV$ on taking expectations,

we get $E(X)=a+hE(U)$, $\qquad E(Y)=b+kE(V)$

$X-E(X)=h\left[U-E(U)\right]$ and $Y-E(Y)=k\left[V-E(V)\right]$

$$\Rightarrow \text{Cov}\,(X,Y)=E\left\{\left[X-E(X)\right]\left[Y-E(Y)\right]\right\}$$

$$=E\left\{h\left[U-E(U)\right]k\left[V-E(V)\right]\right\}=hkE\left\{\left[U-E(U)\right]\left[V-E(V)\right]\right\}$$

$$=hk\ \text{Cov}\,(U,V)\rightarrow(1)$$

$$\sigma_x^2 = E\{X - E(X)\}^2 = E\left\{h^2\left[U - E(U)\right]^2\right\} = h^2 E\{U - E(U)\}^2$$

$$= h^2\sigma_u^2$$

$$\Rightarrow \sigma_x = h\sigma_U \;(h > 0) \rightarrow (2)$$

and similarly $\sigma_y^2 = k^2\sigma_v^2$

$$\Rightarrow \sigma_y = k\sigma_v \;(k > 0) \rightarrow (3)$$

Substituting (1), (2) and (3) in the following

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_x\sigma_y} = \frac{hk\ Cov\ (U,V)}{h\sigma_U k\sigma_V} = \frac{Cov\ (U,V)}{\sigma_U\sigma_V}$$

$$= r(U,V)$$

$$\therefore\ r(X,Y) = r(U,V)$$

(iii) Two independent variables are uncorrelated.

Proof:

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

$$r(X,Y) = \frac{Cov\ (X,Y)}{\sigma_X\sigma_Y} = \frac{0}{\sigma_X\sigma_Y} = 0.$$

Hence two independent variables are uncorrelated.

Note: Two uncorrelated variables need not be independent

For example:

X	-2	-1	1	2	$\sum X = 0$
Y	4	1	1	4	$\sum Y = 10$
XY	-8	-1	1	8	$\sum XY = 0$

$$\overline{X} = \frac{1}{n}\sum X = 0;\ \overline{Y} = \frac{1}{n}\sum Y = \frac{10}{4} = 2.5$$

$$\text{Cov}\left(X,\ Y\right)=\frac{1}{n}\sum XY-\overline{X}\ \overline{Y}=0-0\times 2.5=0$$

$$r\left(X,Y\right)=\frac{\text{Cov}\left(X,Y\right)}{\sigma_X\sigma_Y}=\frac{0}{\sigma_X\sigma_Y}=0\cdot$$

Thus in the above example, the variables X and Y are uncorrelated, but they are not independent, they are related as $Y=X^2$. Hence two uncorrelated variables need not be independent. Reason for this is absence of any linear relationship between the variables X and Y. There may be some other form of relationship between them e.g. quadratic, cubic or trignometric.

Example: Calculate the correlation coefficient for the following X and Y.

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

Solution: Calculation for correlation coefficient

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
544 $\sum X=$	552 $\sum Y=$	37028 $\sum X^2=$	38132 $\sum Y^2=$	37560 $\sum XY=$

$$\overline{X}=\frac{1}{n}\sum X=\frac{544}{8}=68\ ;\ \overline{Y}=\frac{1}{n}\sum Y=\frac{552}{8}=69$$

$$\text{Cov}\left(X,Y\right)=\frac{1}{n}\sum XY-\overline{X}\ \overline{Y}=\frac{37560}{8}-68\times 69$$

$$=4695-4692=3$$

$$\sigma_X = \sqrt{\frac{1}{n}\sum X^2 - \left(\overline{X}\right)^2} = \sqrt{\frac{37028}{8} - (68)^2} = \sqrt{4.5}$$

$$\sigma_Y = \sqrt{\frac{1}{n}\sum Y^2 - \left(\overline{Y}\right)^2} = \sqrt{\frac{38132}{8} - (69)^2} = \sqrt{5.5}$$

$$r(X,Y) = \frac{\text{Cov}\,(X,Y)}{\sigma_X \sigma_Y} = \frac{3}{\sqrt{4.5}\,\sqrt{5.5}} = 0.603$$

Change of Origin and Scale method for the above example:

X	Y	U = X - 68	V = Y - 69	U^2	V^2	UV
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	65	-1	-4	1	16	4
67	68	-1	-1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
		0 $\sum U =$	0 $\sum V =$	36 $\sum U^2 =$	44 $\sum V^2 =$	24 $\sum UV =$

$$\overline{U} = \frac{1}{8} \times 0 = 0\,, \qquad \overline{V} = \frac{1}{8} \times 0 = 0$$

$$\text{Cov}\,(U,V) = \frac{1}{n}\sum UV - \overline{U}\;\overline{V} = \frac{1}{8} \times 24 - 0 \times 0 = 3$$

$$\sigma_U^2 = \frac{1}{n}\sum U^2 - \left(\overline{U}\right)^2 = \frac{1}{8} \times 36 - 0 = 4.5$$

$$\sigma_V^2 = \frac{1}{n}\sum V^2 - \left(\overline{V}\right)^2 = \frac{1}{8} \times 44 - 0 = 5.5$$

$$r(U,\,V) = \frac{\text{Cov}\,(U,V)}{\sigma_U \cdot \sigma_V} = \frac{3}{\sqrt{4.5 \times 5.5}} = 0.603 = r(X,Y)$$

3.3 Partial Correlation:

Correlation between any two variables studied partially i.e. studied after eliminating the linear effect of others from them is called partial correlation. In fact, it remains the partial relationship only between two variables when the effect of other variable/ variables is excluded. Here, we shall consider the partial correlation between two variables by eliminating the linear effect of the third one.

Partial Correlation Coefficient:

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

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{32}^2\right)}}$$

and $$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{\left(1-r_{21}^2\right)\left(1-r_{31}^2\right)}}$$

r_{12}, r_{13}, r_{23} the correlation between X_1 and X_2; X_1 and X_3; X_2 and X_3 respectively.

Example 2: From the following data find the partial correlation coefficient $r_{12.3}$.

$$r_{12} = 0.77; r_{13} = 0.72; r_{23} = 0.52$$

Solution: Partial Correlation Coefficient

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

$$= \frac{0.77 - 0.72 \times 0.52}{\sqrt{\left[1-(0.72)^2\right]\left[1-(0.52)^2\right]}} = 0.62$$

3.4 Multiple Correlation:

Whenever we are interested in studying the joint effect of a group of variables upon a variable not included in the group, our study is that of multiple correlation.

For example, the yield of crop per acre say X_1 depends upon quality of seed X_2, fertility of soil X_3, fertilizer used X_4, irrigation facilities X_5, weather conditions X_6 and so on.

Coefficient of Multiple Correlation:

In a tri-variate distribution in which each of the variables X_1, X_2 and X_3 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. In other words $R_{1.23}$ is the correlation coefficient between X_1 and its estimated value as given by the plane of regression of X_1 on X_2 and X_3. By definition

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

This formula expresses the multiple correlation coefficient interms of the total correlation coefficients between the pairs of variables.

Similarly $R_{2.13}^2$ and $R_{3.12}^2$ are given by

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

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

Properties of Multiple Correlation Coefficient:

1. Multiple correlation coefficient measures the closeness of the association between the observed values and the expected values of a variable obtained from the multiple linear regression of that variable on other variables.

2. Multiple correlation coefficient between observed values and expected values, when the expected values are calculated from a linear relation of the variables determined by the method of least squeres is always greater than that where expected values are calculated from any other linear combination of the variables.

3. $0 \le R_{1.23} \le 1$

4. If $R_{1.23} = 1$, then association is perfect.

5. If $R_{1.23} = 0$, then total and partial correlations involving X_1 are zero.

6. $R_{1.23} \ge r_{12} \cdot r_{13} \cdot r_{23}$

Example 3: Calculate $R_{1.23}, R_{3.12}$ and $R_{2.13}$ for the following data, given

$r_{12} = 0.6; r_{13} = 0.7 ; r_{23} = 0.65$.

Solution:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.6)^2 + (0.7)^2 - 2(0.6)(0.7)(0.65)}{1-(0.65)^2}} = 0.726$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}} = \sqrt{\frac{(0.7)^2 + (0.65)^2 - 2(0.6)(0.7)(0.65)}{1-(0.6)^2}} = 0.757$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} = \sqrt{\frac{(0.6)^2 + (0.65)^2 - 2(0.6)(0.7)(0.65)}{1-(0.7)^2}} = 0.681$$

3.5 Spearman's Rank Correlation Coefficient:

Spearman's rank correlation coefficient is denoted by $'\rho'$ defined as

$$\rho = 1 - \left[\frac{6\sum d_i^2}{n\left(n^2 - 1\right)}\right] \quad \text{where } di = x_i - y_i \text{ ,}$$

n = Number of paired observations.

Proof:

Let (x_i, y_i), $i=1,2,.........,n$ be the ranks of i^{th} individual in two characteristics A and B respectively. Pearson coefficient of correclation between the ranks of X's and Y's is called the rank correlation coefficient between A and B for that group of individuals.

Assuming that no two individuals are bracketed equal in either classification each of the variables X and Y takes values 1,2,..........., n.

Hence $\bar{x} = \bar{y} = \frac{1}{n}(1 + 2 + ... + n) = \frac{1}{n}\ \frac{n(n+1)}{2} = \frac{n+1}{2}$

$$\sigma_x^2 = \frac{1}{n}\sum x_i^2 - \left(\bar{x}\right)^2 = \frac{1}{n}\left[1^2 + 2^2 + ... + n^2\right] - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{1}{n}\ \frac{n(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n+1}{2}\left(\frac{2n+1}{3} - \frac{n+1}{2}\right) = \frac{n^2-1}{12} = \sigma_y^2$$

In general $x_i \neq y_i$; let $d_i = x_i - y_i$

$$d_i = \left(x_i - \overline{x}\right) - \left(y_i - \overline{y}\right) \qquad \left(\because\ \overline{x} = \overline{y}\right)$$

Squaring and summing over 'i' from 1 to n, we get

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}\left[\left(x_i - \overline{x}\right) - \left(y_i - \overline{y}\right)\right]^2$$

dividing both sides by 'n'

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 - 2\frac{1}{n}\sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

$$= \sigma_x^2 + \sigma_y^2 - 2\ \text{Cov}\,(X, Y)$$

$$= 2\sigma_x^2 - \text{Cov}\,(X, Y) \qquad \left(\because\ \sigma_X^2 = \sigma_Y^2\right)$$

$$\frac{1}{n}\sum d_i^2 = 2\sigma_x^2 - 2r\,\sigma_x^2 \qquad \left(\because\ r = \frac{\text{Cov}\,(X,Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X,Y)}{\sigma_X^2} \qquad \because \sigma_X = \sigma_Y\right)$$

$$= 2\sigma_x^2\,(1 - r)$$

$$\therefore\ 1 - r = \frac{\frac{1}{n}\sum d_i^2}{2\sigma_x^2} = \frac{\frac{1}{n}\sum d_i^2}{\frac{2\left(n^2-1\right)}{12}} = \frac{6\sum d_i^2}{n\left(n^2-1\right)}$$

$$\Rightarrow r = 1 - \left(\frac{6\sum d_i^2}{n\left(n^2-1\right)}\right)$$

which is the formula for Spearman's rank correlation coefficient, it is denoted by ρ

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

Repeated Ranks: If any two or more individuals are bracketed equal in any classification with respect to characteristics A or B, or if there is more than one item with the same value in the series. The spearman's formula for calculating the rank correlation coefficient breaks down. In this case common ranks are given to the repeated items. The common rank is the average of the rank which these items would have assumed. If they were slight different from each other and the next item will get the rank next to the ranks already assumed.

In this formula we add the factor $\frac{m\left(m^2 - 1\right)}{12}$ to $\sum d_i^2$ where 'm' is the number of times an item is repeated. The correction factor is to be added for each repeated value.

Limits for rank correlation coefficient (ρ):

$$-1 \le \ \rho \ \le 1$$

Proof: 'ρ' **is maximum:** If $\sum d_i^2$ is minimum i.e. if each of the deviation d_i is minimum, but minimum value of d_i is zero i.e. $x_i = y_i$. Hence the maximum value of ρ is +1 i.e. $\rho \le 1$.

'ρ' **Minimum:** If $\sum d_i^2$ is maximum which will be so if the ranks X and Y are in opposite directions as given below

X	1	2	....	n-1	n
Y	n	n-1	....	2	1

This gives us $x_i + y_i = n + 1 \ \forall\ i = 1, 2, \ldots\ldots\ldots, n$

$$x_i - y_i = d_i$$

Also $2x_i = n + 1 + d_i \Rightarrow d_i = 2x_i - (n+1)$

$$\sum d_i^2 = \sum\left(4x_i^2 + (n+1)^2 - 2(n+1)\cdot 2x_i\right)$$

$$= 4\sum x_i^2 + n(n+1)^2 - 4(n+1)\sum x_i$$

$$=\frac{4n(n+1)(2n+1)}{6}+n(n+1)^2-4(n+1)\frac{n(n+1)}{2}$$

$$\sum d_i^2=\frac{n\left(n^2-1\right)}{3}$$

but $\rho=1-\dfrac{6\sum d_i^2}{n\left(n^2-1\right)}$

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

i.e. $-1\leq\rho$

$\therefore \quad -1\leq\rho\leq 1$

3.6 Problems:

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

X	10	12	18	24	23	27
Y	13	18	12	25	30	10

Solution:

X	Y	XY	X^2	Y^2
10	13	130	100	169
12	18	216	144	324
18	12	216	324	144
24	25	600	576	625
23	30	690	529	900
27	10	270	729	100
114 $\sum X=$	108 $\sum Y=$	2122 $\sum XY=$	2402 $\sum X^2=$	2262 $\sum Y^2=$

n = 6; $\sum X = 114$; $\sum Y = 108$; $\sum XY = 2122$; $\sum X^2 = 2402$; $\sum Y^2 = 2262$

$$\overline{X} = \frac{\sum X}{n} = \frac{114}{6} \; ; \overline{Y} = \frac{\sum Y}{n} = \frac{108}{6}$$

Karl Pearson Coefficient of Correlation $(r) = \dfrac{\frac{1}{n}\sum XY - \overline{X}\,\overline{Y}}{\sqrt{\frac{1}{n}\sum X^2 - (\overline{X})^2}\sqrt{\frac{1}{n}\sum Y^2 - (\overline{Y})^2}}$

$$= \frac{\frac{1}{6}\times 2122 - \frac{114}{6}\times\frac{108}{6}}{\sqrt{\frac{1}{6}\times 2402 - \left(\frac{114}{6}\right)^2}\sqrt{\frac{1}{6}\times 2262 - \left(\frac{108}{6}\right)^2}} = 0.255$$

2. Given the following frequency distribution, find the correlation coefficient between X and Y.

Y \ X	5	10	Total
10	30	20	50
20	20	30	50
Total	50	50	100

Solution: Karl Pearson coefficient of correlation $(r) = \dfrac{n\sum f_{XY} - (\sum fx)(\sum fy)}{\sqrt{n\sum fx^2 - (\sum fx)^2}\sqrt{n\sum fy^2 - (\sum fy)^2}}$

Y \ X	5	10	f	fy	fy^2	fxy
10	30	20	50	500	5000	3500
20	20	30	50	1000	20,000	8000
f	50	50	100 $= N$	1500 $= \sum fy$	25,000 $= \sum fy^2$	11,500 $= \sum fxy$
fx	250	500	750 $= \sum fx$			
fx^2	1250	5000	6250 $= \sum fx^2$			
fxy	3500	8000	11500 $= \sum fxy$			

$$r = \frac{100 \times 11500 - 750 \times 1500}{\sqrt{100 \times 6250 - (750)^2}\ \sqrt{100 \times 25,000 - (1500)^2}} = 0.2$$

3. Calculate rank correlation coefficient for the following data.

X	52	63	45	36	72	65	47	25
Y	62	53	51	25	79	43	60	33

Solution:

X	Y	Rank of X R_1	Rank of Y R_2	$d_i = \lvert R_1 - R_2 \rvert$	d_i^2
52	62	4	2	2	4
63	53	3	4	1	1
45	51	6	5	1	1
36	25	7	8	1	1
72	79	1	1	0	0
65	43	2	6	4	16
47	60	5	3	2	4
25	33	8	7	1	1
				$\sum d_i^2 =$	28

n = 8

Rank correlation coefficient $(\rho) = 1 - \left(\frac{6 \sum d_i^2}{n\left(n^2 - 1\right)} \right)$

$$= 1 - \left(\frac{6 \times 28}{8\left(8^2 - 1\right)} \right) = 1 - 0.33 = 0.67$$

4. Compute rank correlation coefficient from the following data.

X	60	15	20	28	12	40	80	20
Y	10	40	30	50	30	20	60	30

Solution:

X	Y	Rank of X R_1	Rank of Y R_2	$d_i = \lvert R_1 - R_2 \rvert$	d_i^2
60	10	2	8	6	36
15	40	7	3	4	16
20	30	5.5	5	0.5	0.25
28	50	4	2	2	4
12	30	8	5	3	9
40	20	3	7	4	16
80	60	1	1	0	0
20	30	5.5	5	0.5	0.25

$\sum d_i^2 =$ 81.50

In 'X' - Series:

$$m_1 = 2;\ C \cdot F_1 = \frac{m_1\left(m_1^2 - 1\right)}{12} = \frac{2(4-1)}{12} = 0.5$$

In 'Y' Series:

$$m_2 = 3\,;\ C.F_2 = \frac{3(9-1)}{12} = 2$$

n = 8; $\sum d_i^2 = 81.5$

Rank correlation coefficient $$(\rho) = 1 - \left(\frac{6\left(\sum d_i^2 + C.F_1 + C \cdot F_2 + \text{..........}\right)}{n\left(n^2 - 1\right)}\right)$$

$$= 1 - \left[\frac{6(81.5 + 0.5 + 2)}{8\left(8^2 - 1\right)}\right]$$

$$= 1 - \frac{6 \times 84}{8 \times 63}$$

= 1 - 1 =0

5. In a trivariate distribution it is found that

$$r_{12} = 0.7, \qquad r_{13} = 0.61, \qquad r_{23} = 0.4$$

Find the value of $r_{23.1}$, $r_{13.2}$ and $r_{12.3}$

Solution:

$$r_{23.1} = \frac{r_{23} - r_{12}\,r_{13}}{\sqrt{1-r_{12}^2}\,\sqrt{1-r_{13}^2}} = \frac{0.4 - 0.7 \times 0.61}{\sqrt{1-(0.7)^2}\,\sqrt{1-(0.61)^2}} = -0.418$$

$$r_{13.2} = \frac{r_{13} - r_{12}\,r_{23}}{\sqrt{1-r_{12}^2}\,\sqrt{1-r_{23}^2}} = \frac{0.61-(0.7)(0.4)}{\sqrt{1-(0.7)^2}\,\sqrt{1-(0.4)^2}} = 0.504$$

$$r_{12.3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{1-r_{13}^2}\,\sqrt{1-r_{23}^2}} = \frac{0.7-(0.61)(0.4)}{\sqrt{1-(0.61)^2}\,\sqrt{1-(0.4)^2}} = 0.628$$

6. If $r_{12} = 0.9, r_{13} = 0.75, r_{23} = 0.7$ Find $R_{1.23}$

Solution:

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

$$= \frac{(0.9)^2 + (0.75)^2 - 2(0.9)(0.75)(0.7)}{1-(0.7)^2} = 0.838$$

$$R_{1.23} = \sqrt{0.838} = 0.916$$

3.7 Model Questions and Exercises:

1. Define Karl Pearson's coefficient of correlation and show that correlation coefficient is unaltered by change of origin and scale.

2. Derive the limits for Karl Pearson coefficient of correlation.

3. Calculate Karl Pearson coefficient of correlation for the following data.

X	23	27	28	29	30	31	33	35	36	39
Y	18	22	23	24	25	26	28	22	30	32

4. Derive the Spearman's rank correlation coefficient.

5. Derive the limits for Spearman's rank correlation coefficient.

3.8 Summary:

Concepts of correlation coefficient and its properties, rank correlation coefficient, tied and untied ranks, multiple and partial correlation coefficients are discussed.

Some problems are solved to illustrate the concepts. Some model questions and a problem are given to the students to prepare on their own.

3.9 Technical Terms:

Correlation

Correlation Coefficient

Rank Correlation Coefficient

Partial Correlation

Multiple correlation.

Lesson Writer

A. Mohan Rao

Lesson 4

REGRESSION

Objective:

After studying the lesson the students will have clear comprehension in the concepts and applications of simple linear regression, comparison between correlation and regression, coefficient of determination, correlation ratio.

Structure of the Lesson:

4.1 Introduction:

The term "regression" literally means "stepping back" towards the average. It was first used by a British biometrician Sir Francis Galton (1822 - 1911) in connection with the inheritance of stature. Galton found that the off springs of abnormally tall or short parents tend to "regress" or "step back" to the average population height. But the term "regression" as now used in statistics is only a convenient term without having any reference to biometry.

In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called dependent variable and the variable which influences the values or is used for prediction, is called independent variable. In regression analysis independent variable is also known as regressor or predictor or explanatory variable while the dependent variable is also known as regressed or explained variable.

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

4.2 Simple Linear Regression and Properties of Regression Coefficients:

If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster round some curve called the "curve of regression". If the curve is a straight line, it is called the line of regression and there is said to be simple linear regression between the variables, otherwise regression is said to be curvilinear.

The line of regression is the line which gives the best eslimate to the value of one variable for any specific value of the other variable. The line of regression is the line of best fit obtained by the principle of least squares.

Let us suppose that in the bivariate distribution $(x_i, y_i);\ i=1,2,.........,n$; X is independent variable and Y is the dependent variable.

Let the regression of Y on X be $Y = a + bX \rightarrow (1)$

According to the principle of least squares, the normal equations for estimating a and b are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \rightarrow (2)$$

and $$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \rightarrow (3)$$

Dividing (2) by n, we get

$$\bar{y} = a + b\,\bar{x} \rightarrow (4)$$

Thus the line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$.

Now $$\mu_{11} = \text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \text{Cov}(X,Y) + \bar{x}\,\bar{y} \rightarrow (5)$$

Also $$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - \bar{x}^2 \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma_x^2 + (\bar{x})^2 \rightarrow (6)$$

Dividing (3) by n $$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i y_i = a\frac{1}{n}\sum_{i=1}^{n} x_i + b\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^2 \rightarrow (7)$$

Substituting (5) and (6) in (7) we get

$$\text{Cov}(X,Y) + \bar{x}\,\bar{y} = a\bar{x} + b\left(\sigma_x^2 + (\bar{x})^2\right) \rightarrow (8)$$

$(4) \times \bar{x} \Rightarrow$ $$\bar{x}\,\bar{y} = a\bar{x} + b(\bar{x})^2 \rightarrow (9)$$

$(8) - (9) \Rightarrow$ $$\text{Cov}(X,Y) = b\sigma_x^2 \Rightarrow b = \frac{\text{Cov}(X,Y)}{\sigma_x^2}.$$

Since 'b' is the slope of the line of regression of Y on X and since the line of regression passes through the point $\left(\bar{x}, \bar{y}\right)$; its equation is

$$\left(Y-\overline{Y}\right)=b\left(X-\overline{X}\right)$$

$$=\frac{\operatorname{Cov}(X, Y)}{\sigma_x^2}\left(X-\overline{X}\right)$$

$$\Rightarrow\left(Y-\overline{Y}\right)=r\,\frac{\sigma_Y}{\sigma_X}\left(X-\overline{X}\right) \qquad \left(\because \operatorname{Cov}(X, Y)=r\sigma_X\cdot\sigma_Y\right)$$

which is the regression equation of Y on X.

Similarly starting with $X=A+BY$, continuing in the above manner we get

$$X-\overline{X}=r\frac{\sigma_X}{\sigma_Y}\left(Y-\overline{Y}\right)$$

which is the regression equation of X on Y.

Properties of Regression Coefficients:

Regression coefficient Y on X $\left(b_{YX}\right)=\dfrac{\operatorname{Cov}(X,Y)}{\sigma_x^2}=r\dfrac{\sigma_Y}{\sigma_X}$

Regression coefficient of X on Y $\left(b_{XY}\right)=\dfrac{\operatorname{Cov}\,(X,Y)}{\sigma_y^2}=r\dfrac{\sigma_X}{\sigma_Y}$

1. Correlation coefficient is the geometric mean between the regression coefficients.

Proof: $b_{XY}\cdot b_{YX}=r\dfrac{\sigma_X}{\sigma_Y}\cdot r\dfrac{\sigma_Y}{\sigma_X}=r^2$

$$\therefore r=\pm\sqrt{b_{XY}\cdot b_{YX}}$$

Note: If regression coefficients are positive 'r' is positive and if regression coefficients are negative 'r' is negative.

2. If one of the regression coefficients is greater than unity, the other must be less than unity.

Proof: Let $b_{YX}>1$, then we have to show that $b_{XY}<1$

Now $b_{YX} > 1 \Rightarrow \dfrac{1}{b_{YX}} < 1$

Also $r^2 \le 1 \Rightarrow b_{YX} \cdot b_{XY} \le 1$

Hence $b_{XY} \le \dfrac{1}{b_{YX}} < 1$

$\therefore \quad b_{XY} < 1$

3. Arithmetic mean of the regression coefficients is greater than the correlation coefficient r, provided $r > 0$.

Proof: We have to prove that $\frac{1}{2}\left(b_{YX} + b_{XY}\right) \ge r$

or $\dfrac{1}{2}\left(r\dfrac{\sigma_Y}{\sigma_X} + r\dfrac{\sigma_X}{\sigma_Y}\right) \ge r$ or $\dfrac{\sigma_Y}{\sigma_X} + \dfrac{\sigma_X}{\sigma_Y} \ge 2 \qquad (\because r>0)$

$\Rightarrow \sigma_y^2 + \sigma_x^2 - 2\sigma_X\sigma_Y \ge 0$ i.e. $(\sigma_Y - \sigma_X)^2 \ge 0$

which is always true, since the square of real quantity is ≥ 0.

4. Regression coefficients are independent of change of origin but not of scale.

Proof: Let $U = \dfrac{X-a}{h}$; $V = \dfrac{Y-b}{k} \Rightarrow X = a + bU$; $Y = b + kV$

where a, b, h (> 0) and k (>0) are constants.

Then $\mathrm{Cov}(X, Y) = hk\ \mathrm{Cov}(U,V)$; $\sigma_x^2 = h^2\ \sigma_U^2$ and $\sigma_Y^2 = k^2\sigma_V^2$

$$b_{YX} = \frac{\mathrm{Cov}(X,Y)}{\sigma_x^2} = \frac{hk\ \mathrm{Cov}(U,V)}{h^2\ \sigma_U^2} = \frac{k}{h}\, b_{VU}$$

Similarly, we can prove that

$b_{XY} = \left(h/k\right) b_{UV}$.

4.3 Angle between two regression lines:

Equations of the lines of regression of Y on X, and X on Y are

$$Y-\overline{Y}=r\frac{\sigma_Y}{\sigma_X}\left(X-\overline{X}\right) \text{ and } X-\overline{X}=r\frac{\sigma_X}{\sigma_Y}\left(Y-\overline{Y}\right)$$

Slopes of these lines are $r\frac{\sigma_Y}{\sigma_X}$ and $\frac{\sigma_Y}{r\sigma_X}$ respectively.

If 'θ' is the anlge between the two lines of regression

then $$\tan\theta=\frac{r\frac{\sigma_Y}{\sigma_X}\sim\frac{\sigma_Y}{r\sigma_X}}{1+r\frac{\sigma_Y}{\sigma_X}\cdot\frac{\sigma_Y}{r\sigma_X}}=\frac{r^2\sim 1}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}\right)$$

$$=\frac{1-r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}\right) \qquad \left(\because\ r^2\le 1\right)$$

$$\therefore\ \theta=\tan^{-1}\left\{\frac{1-r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}\right)\right\}$$

Case (i): If $r=0, \tan\theta=\infty\Rightarrow\theta=\pi/2$

The lines of regression become perpendicular to each other.

Case (ii): If $r=\pm 1, \tan\theta=0 \Rightarrow \theta=0 \text{ or } \pi$

lines of regression either coincide or they are parallel to each other, but they pass through $\left(\overline{x},\overline{y}\right)$, hence they coincide.

Example 1: Find the most likely price in Bombay corresponding to the price of Rs 70 at Kolkatta from the following data.

	Kolkatta	**Bombay**
Mean Price	65	67
S.D	2.5	3.5

Correlation coefficient between the prices of commodities in the two cities is 0.8

Solution: Let the prices, (in Rs). in Bombay and Kolkatta be denoted by Y and X respectively. Then we are given

$$\overline{X} = 65;\ \overline{Y} = 67;\ \sigma_X = 2.5;\ \sigma_Y = 3.5 \text{ and } r(X, Y) = 0.8$$

Line of regression of Y on X is

$$Y - \overline{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right)$$

$$Y = \overline{Y} + r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right)$$

$$= 67 + 0.8\left(\frac{3.5}{2.5}\right)(X - 65)$$

When $X = 70;\ \hat{Y} = 67 + 0.8\left(\frac{3.5}{2.5}\right)(70 - 65)$

$$= 72.6$$

Example 2: The following results were worked out from scores in Mathematics and English in a certain examination.

	Scores in Mathematics (X)	Scores in English (Y)
Mean	39.5	47.5
S.D	10.8	17.8

Correlation coefficient between X and Y is 0.42

Find both the regression lines. Using these regressions estimate the value of Y for X = 50 and estimate the value of X for Y = 30.

Solution: Given $\overline{X} = 39.5$; $\overline{Y} = 47.5;\ \sigma_X = 10.8;\ \sigma_Y = 17.8;\ r(X, Y) = 0.42$

Regression equation of Y on X is

$$Y - \overline{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right)$$

$$Y = \overline{Y} + r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right) = 47.5 + 0.42\left(\frac{17.8}{10.8}\right)(X - 39.5)$$

$$= 47.5 + 0.69\ X - 0.69\ X\ 39.5$$

$$Y = 0.69\ X + 2.25$$

Regression equation of X on Y is

$$X = \overline{X} + r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right)$$

$$X = 39.5 + 0.42\left(\frac{10.8}{17.8}\right)\left(Y - 47.5\right)$$

$$= 39.5 + 0.25Y - 0.25 \times 47.5$$

$$= 39.5 + 0.25Y - 11.875$$

$$X = 0.25Y + 27.63$$

If $X = 50;\ \hat{Y} = 0.69(50) + 20.25$

$$= 54.75$$

If $Y = 30\ ;\ \hat{X} = 0.25(30) + 27.63$

$$= 35.13$$

Example 3: In a partially destroyed laboratory record of an analysis of correlation data the following results only are legible.

$$\sigma_X^2 = 9;\ 8X - 10Y + 66 = 0;\ 40X - 18Y = 214$$

what were (a) The mean values of X and Y

(b) σ_Y (c) $r(X,Y)$

Solution: (a) Solving regression equations we get X & Y as follows:

$$8X - 10Y + 66 = 0 \rightarrow (1)$$

$$40X - 18Y - 214 = 0 \rightarrow (2)$$

$(1) \times 5 \Rightarrow$ $40X - 50Y + 330 = 0$

$(-)$

$$32Y - 544 = 0$$

$$\Rightarrow Y = \frac{544}{32} = 17$$

Substituting Y = 17 in (1), $X = \frac{10\times 17 - 66}{8} = 13$

Since regression equations pass through $\left(\overline{X}, \overline{Y}\right)$

$$\therefore\ \overline{X} = 13;\ \overline{Y} = 17.$$

(b) We have $b_{YX} = r\frac{\sigma_Y}{\sigma_X} \Rightarrow \sigma_Y = b_{YX}\cdot\frac{\sigma_X}{r} = \frac{8}{10}\times\frac{3}{0.6} = 4$

(c) From the regression equations

Let $8X - 10Y + 66 = 0$ be the Y on X $\Rightarrow Y = \frac{8}{10}X + \frac{66}{10}$

$40X - 18Y = 214$ be the X on Y $\Rightarrow X = \frac{18}{40}\ Y + \frac{214}{40}$

$$\therefore\ b_{YX} = \frac{8}{10}$$

$$b_{XY} = \frac{18}{40}$$

$$r^2 = b_{YX}\cdot b_{XY} = \frac{8}{10}\times\frac{18}{40} = 0.36$$

$$r = \sqrt{0.36} = \pm\ 0.6$$

4.4 Comparison of Correlation and Regression:

Correlation	Regression
1. Literally means relationship between two or more variables.	1. Regression means stepping back towards the average value and is a mathematical measure expressing the average relationship between the two variables.
2. Correlation coefficient is independent of change of origin and scale.	2. Regression coefficients are independent of change of origin but not scale.
3. There may be non sense correlation between two variables	3. No such type of non sense regression between two variables
4. The coefficient of correlation is the measure of direction and degree of linear relationship between two variables, so it is symmetric i.e. $r_{XY} = r_{YX}$	4. Its analysis deals with the functional relationship between the two variables. One is independent and the other is dependent, hence it is not symmetric i.e. $b_{XY} \neq b_{YX}$
5. Correlation has limited applications since it is confined to the study of linear relationship between the variables.	5. Regnession has much wider applications as it studies linear and non-linear relationship between the variables.

4.5 Coefficient of Determination:

$\sum_i \left(\hat{y}_i - \overline{Y}\right)^2$ is the sum of the squares of the errors of the predicted values $\hat{y}_i$ from the mean value $\overline{Y}$, is called explained variation.

S_y^2 is called the total variation.

Coefficient of determination is the ratio of explained variation to the total variation.

i.e. Coefficient of determination $= \dfrac{\sum_i \left(\hat{y}_i - \overline{Y}\right)^2}{S_Y^2}$

This gives a measure of the usefulness of the line of regression in prediction from x.

Coefficient of determination lies between 0 and 1.

Unexplained Variation $D = \sum_i (y_i - \hat{y}_i)^2$.

Coefficient of determination can also expressed as

$$\frac{\sum_i (\hat{y}_i - \overline{Y})^2}{S_Y^2} = 1 - \frac{D}{S_Y^2} .$$

Example 4: For the following straight line data calculate coefficient of determination.

X	0	1	2	3	4	5
Y	1	1.8	3.3	4.5	6.3	10

Solution: Let $Y = a + bx \rightarrow (1)$ be the straight line and normal equations are

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

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

	X	Y	XY	X^2	Y^2
	0	1	0	0	1.00
	1	1.8	1.8	1	3.24
	2	3.3	6.6	4	10.89
	3	4.5	13.5	9	20.25
	4	6.3	25.2	16	39.69
	5	10	50.0	25	100.00
Total	$\sum X = 15$	$\sum Y = 26.9$	$\sum XY = 97.1$	$\sum X^2 = 55$	$\sum Y^2 = 175.07$

Substituting these values in normal equations.

$$26.9 = 6a + 15b \rightarrow (2)$$

$$97.1 = 15a + 55b \rightarrow (3)$$

$$(2)\times 5 \Rightarrow \; 134.5 = 30a + 75b$$

$$(3)\times 2 \Rightarrow 194.2 = 30a + 110b$$

$$+59.7 = +35b$$

$$\Rightarrow \; b = \frac{59.7}{35} = 1.7$$

Substituting b = 1.7 in (2)

we get a = 0.23

$$\boxed{\therefore \; Y = 0.23 + 1.7X}$$

$$\overline{Y} = \frac{\sum Y}{n} = \frac{26.9}{6}$$

= 4.48

$$S_Y^2 = \sum Y^2 - n\left(\overline{Y}\right)^2$$

$$= 175.07 - 6(4.48)^2 = 54.65$$

By substituting the different values of X in the above straight line we get $\hat{y}_i$ values as follows:

X	Y	$\hat{y}_i$	$\left(\hat{y}_i - \overline{Y}\right)^2$
0	1	0.23	18.0625
1	1.8	1.93	6.5025
2	3.3	3.63	0.7225
3	4.5	5.33	0.7225
4	6.3	7.03	6.5025
5	10	8.73	18.0625
			50.5750

$$\sum_i \left(\hat{y}_i - \overline{Y}\right)^2 = 50.575$$

$$S_Y^2 = 54.65$$

Coefficient of determination $= \dfrac{50.575}{54.65} = 0.925$

4.6 Correlation Ratio:

We might come across bivariate distributions where r may be very low or even zero but the regression may be strong, or even perfect. Correlation ratio 'η' is the appropriate measure of curvilinear relationship between the two variables. 'η' measures the concentration of points about the curve of best fit. If regression is linear $\eta = r$, other wise $\eta > r$.

Measure of Correlation Ratio: Suppose corresponding to the values $x_i \ (i=1,2,.........,m)$ of the variable X, the variable Y takes the values Y_{ij}, with respective frequencies $f_{ij,\ j=1,2,.............n}$. A typical arrangement is shown below:

Y \ X	1	2	....	i	....	m	Total
1	f_{11}	f_{21}	....	f_{i1}	....	f_{m1}	
2	f_{12}	f_{22}	....	f_{i2}	....	f_{m2}	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
j	f_{1j}	f_{2j}	....	f_{ij}	....	f_{mj}	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
n	f_{1n}	f_{2n}	....	f_{in}	....	f_{mn}	
$\sum_j f_{ij} = n_i$	n_1	n_2	....	n_i	....	n_m	$\sum_{i=1}^{m} n_i = N$
$\sum_j f_{ij} Y_{ij} = T_i$	T_1	T_2	....	T_i	....	T_m	$\sum_i T_i = T$

Though all the x's in the i^{th} vertical array have the same value, the y's are different. A typical pair of values in the arrays is (x_i, Y_{ij}) with frequency f_{ij}. Thus the first suffix 'i' indicates the vertical array while the second suffix 'j' indicates the positions of 'y' in that array.

Let $\sum_{j=1}^{n} f_{ij} = n_i$ and $\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} f_{ij}\right) = \sum_{i=1}^{m} n_i = N$.

If $\overline{Y_i}$ and $\overline{Y}$ denote the means of the i^{th} array and the overall mean respectively, then

$$\overline{Y_i} = \frac{\sum\limits_{j} f_{ij} Y_{ij}}{\sum\limits_{j=1}^{n} f_{ij}} = \frac{\sum\limits_{j} f_{ij} Y_{ij}}{n_i} = \frac{T_i}{n_i}$$

$$\because \quad \overline{Y} = \frac{\sum\limits_{i}\sum\limits_{j} f_{ij} Y_{ij}}{\sum\limits_{i}\sum\limits_{j} f_{ij}} = \frac{\sum\limits_{i} n_i \overline{Y}_i}{\sum\limits_{i} n_i} = \frac{T}{N}\,.$$

In other words $\overline{Y}$ is the weighted mean of all the array means, the weights being the array frequencies.

Definition: Correlation Ratio: The correlation ratio of Y on X, usually denoted by η_{YX} is given by $\eta_{YX}^2 = 1 - \frac{\sigma_{eY}^2}{\sigma_Y^2}$,

where $\sigma_{eY}^2 = \frac{1}{N}\sum\limits_{i}\sum\limits_{j} f_{ij}\left(Y_{ij} - \overline{Y_i}\right)^2$ and $\sigma_Y^2 = \frac{1}{N}\sum\limits_{i}\sum\limits_{j} f_{ij}\left(Y_{ij} - \overline{Y}\right)^2$

or $\eta_{YX}^2 = \frac{\sigma_{mY}^2}{\sigma_Y^2}$ where $\sigma_{mY}^2 = \sum\limits_{i} n_i \left(\overline{Y}_i - \overline{Y}\right)^2 \Big| N$

$$\text{or} \quad \eta_{YX}^2 = \frac{\left[\sum\limits_{i}\left[\frac{T_i^{\,2}}{n_i}\right] - \frac{T^2}{N}\right]}{N\,\sigma_Y^2}$$

Properties:

(1) η_{YX} lies between -1 and +1.

Proof: By def $\eta_{YX}^2 = 1 - \frac{\sigma_{eY}^2}{\sigma_Y^2}$

$$\Rightarrow \frac{\sigma_{eY}^2}{\sigma_Y^2} = 1 - \eta_{YX}^2$$

Since σ_{eY}^2 and σ_Y^2 are nonnegative. We have

$$1-\eta_{YX}^2 \geq 0 \Rightarrow \eta_{YX}^2 \leq 1$$

$$\Rightarrow |\eta_{YX}| \leq 1$$

$$\Rightarrow -1 \leq \eta_{YX} \leq 1.$$

(2) η_{YX}^2 is independent of change of origin and scale of mesurements.

(3) In general $\eta_{YX} \neq \eta_{XY}$

(4) $|\eta_{YX}| \geq |r|$

Proof: The sum of squares of deviations in any array is minimum when measured from its mean, we have

$$\sum_i \sum_j f_{ij}\left(Y_{ij} - \overline{Y_i}\right)^2 \leq \sum_i \sum_j f_{ij}\left(Y_{ij} - \hat{Y}_{ij}\right)^2 \rightarrow (1)$$

where $\hat{Y}_{ij}$ is the estimate of Y_{ij} for given value of $X = x_i$ (say) as given by the line of regression of Y on X i.e. $\hat{Y}_{ij} = a + bx_i \; (j = 1, 2, ..., n)$.

But $$\sum_i \sum_j f_{ij}\left(Y_{ij} - \overline{Y_i}\right)^2 = N\,\sigma_{eY}^2 = N\sigma_Y^2\left(1-\eta_{YX}^2\right) \rightarrow (2)$$

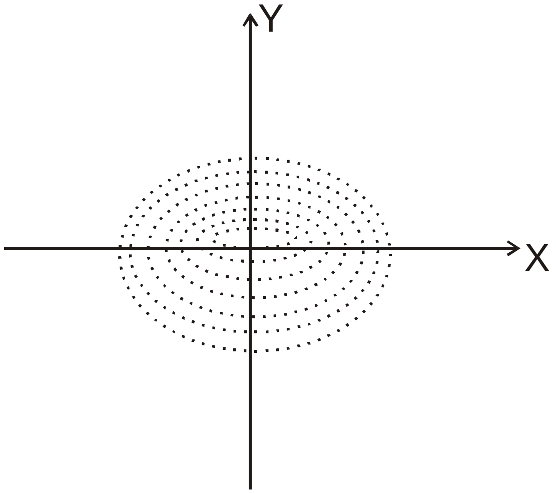
$$\sum_i \sum_j f_{ij}\left(Y_{ij} - a - bx_i\right)^2 = N\sigma_Y^2\left(1-r^2\right) \rightarrow (3)$$

From (2) and (3) in (1) $\Rightarrow 1-\eta_{YX}^2 \leq 1-r^2$

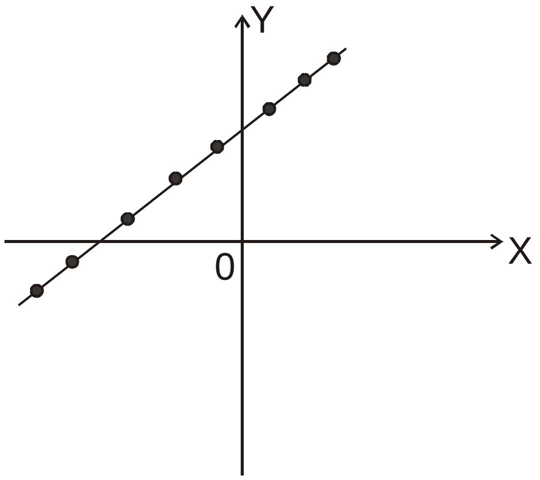
i.e. $\eta_{YX}^2 \geq r^2 \Rightarrow |\eta_{YX}| \geq |r|$

Thus the absolute value of the correlation ratio can never be lessthan absolute value of r.

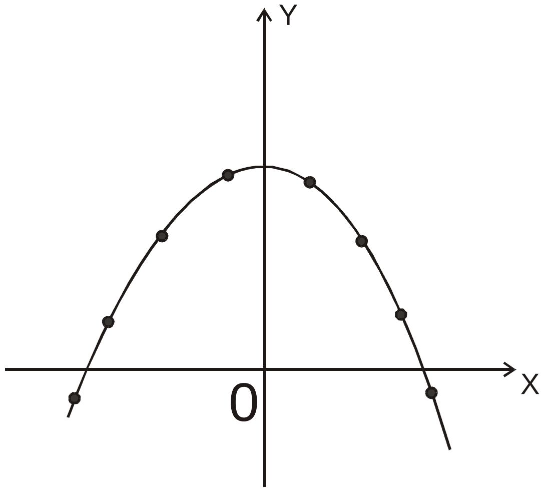
Exhibiting the relationship between 'r' and 'η_{YX}' through diagrams



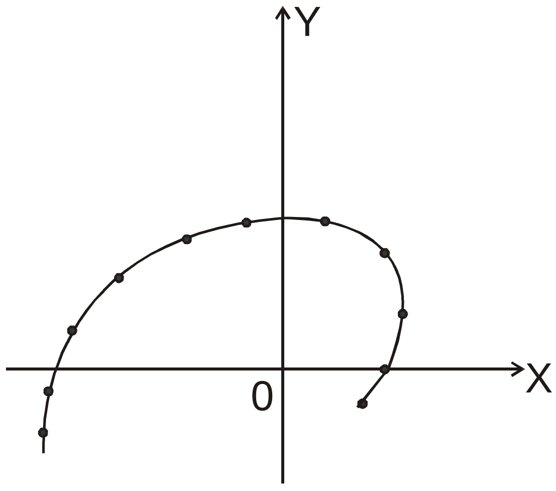
$r = 0 \; ; \eta_{YX} = \eta_{XY} = 0$



$r = 1, \eta_{YX} = \eta_{XY} = 1$



$r = 0, \eta_{YX} = 1, \eta_{XY} = 0$



$\eta_{YX} > r$

4.7 Problems:

Problem 1: Out of the two lines of regression given by $X+2Y-5=0$ and $2X+3Y-8=0$ which one is the regression line of X on Y? Use the equations to find the mean values of X and Y. If the variance of X is 12, calculate the variance of Y.

Solution: $X+2Y-5=0 \Rightarrow 2Y=-X+5$

$$Y=-\frac{X}{2}+5/2$$

$2X+3Y-8=0 \Rightarrow 2X=-3Y+8$

$$X=-\frac{3}{2}Y+\frac{8}{2}$$

Regression equation of Y on X is $X+2Y-5=0$ and regression coefficient is $\boxed{b_{YX}=-\frac{1}{2}}$ Regression equation of X on Y is $2X+3Y-8=0$ and the regression coefficient is $\boxed{b_{XY}=-\frac{3}{2}}$

Since $r^2=b_{YX}\cdot b_{XY}=\frac{-1}{2}\times\frac{-3}{2}=\frac{3}{4}\leq 1$,

which is true.

Solving these two equations we get means of X and Y.

$$X+2Y-5=0 \rightarrow (1)$$

$$\cancel{2X}+3Y-8=0 \rightarrow (2)$$

$$(1)\times 2 \Rightarrow \cancel{2X}+4Y-10=0$$

$$-Y+2=0 \Rightarrow \boxed{Y=2}$$

Substituting Y = 2 in (1) X + 2X2 - 5 = 0

$$\Rightarrow X=1.$$

$\therefore \overline{X}=1$ and $\overline{Y}=2$.

If variance of X is 12; $\sigma_X^2 = 12$

$$b_{YX} = \frac{-1}{2} = r\frac{\sigma_Y}{\sigma_X}$$

$$b_{YX}^2 = \frac{1}{4} = r^2\frac{\sigma_y^2}{\sigma_x^2}$$

$$\Rightarrow \sigma_Y^2 = \frac{1}{4}\frac{\sigma_X^2}{r^2} = \frac{1}{4}\times\frac{12}{3/4} = \frac{1\times12\times4}{4\times3} = 4$$

$$\therefore\ \sigma_Y^2 = 4$$

Problem 2: Form the regression lines of Y on X and X on Y for the following data.

$$\sum X = 70\ ;\ \sum Y = 83\ ;\ \sum X^2 = 590\ ;\ \sum Y^2 = 755\ ;\ \sum XY = 640\ ;\ n = 10$$

Solution: $\overline{X} = \frac{\sum X}{n} = \frac{70}{10} = 7$; $\overline{Y} = \frac{\sum Y}{n} = \frac{83}{10} = 8.3$

$$\sigma_x^2 = \frac{1}{n}\sum X^2 - \left(\overline{X}\right)^2 = \frac{1}{10}\times 590 - (7)^2 = 10\ ;\ \sigma_X = 3.16$$

$$\sigma_Y^2 = \frac{1}{n}\sum Y^2 - \left(\overline{Y}\right)^2 = \frac{1}{10}\times 755 - (8.3)^2 = 6.61\ ;\ \sigma_Y = 2.57$$

$$\text{Cov}(X, Y) = \frac{1}{n}\sum XY - \overline{X}\,\overline{Y} = \frac{1}{10}\times 640 - 7\times 8.3 = 5.9$$

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} = \frac{5.9}{3.16\times 2.57} = 0.73$$

Regression equation of X on Y

$$X = \overline{X} + r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right)$$

$$= 7 + 0.73\left(\frac{3.16}{2.57}\right)(Y - 8.3)$$

$$X = 0.39Y + 3.763$$

Regression equation of Y on X

$$Y = \overline{Y} + r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right)$$

$$= 8.3 + 0.73\left(\frac{2.57}{3.16}\right)(X-7)$$

$$Y = 4.17 + 0.59X$$

Problem 3: If X and Y are two regression lines with equal means and if $Y = aX + b$ and $X = \alpha Y + \beta$ prove that $\frac{b}{\beta} = \frac{1-a}{1-\alpha}$ and also find out the joint mean of two variables.

Solution: $Y = aX + b \rightarrow (1)$

$X = \alpha Y + \beta \rightarrow (2)$

$Y - \overline{Y} = a\left(X - \overline{X}\right)$

$X - \overline{X} = \alpha\left(Y - \overline{Y}\right)$

$\overline{X} = \overline{Y} = m \text{ (say)}$

$Y - m = a(X - m)$

$Y = aX + m(1-a) \rightarrow (3)$

$Y = aX + b \rightarrow (1)$ From (1) & (3) $b = m(1-a)$

$$\Rightarrow m = \frac{b}{1-a} \rightarrow (5)$$

$X = \alpha Y - \alpha m + m = \alpha Y + m(1-\alpha)$

$$X = \alpha Y + m(1-\alpha) \rightarrow (4)$$

From (2) & (4) $\beta = m(1-\alpha)$

$$m = \frac{\beta}{1-\alpha} \rightarrow (6)$$

From (5) and (6) $m = \dfrac{b}{1-a} = \dfrac{\beta}{1-\alpha}$

$$\Rightarrow \frac{b}{\beta} = \frac{1-a}{1-\alpha}$$

$\therefore$ Joint mean of two variables is $\dfrac{b}{1-a} = \dfrac{\beta}{1-\alpha}$

Problem 4: Find two regression lines X on Y and Y on X for the following data.

X	45	53	74	82	38	70
Y	52	38	73	48	74	90

Solution: To get regression lines we must calculate the following values

$$\overline{X} = \frac{\sum X}{n}\ ;\ \overline{Y} = \frac{\sum Y}{n}\ ;\ \sigma_X^2 = \frac{1}{n}\sum X^2 - \left(\overline{X}\right)^2$$

$$\sigma_Y^2 = \frac{1}{n}\sum Y^2 - \left(\overline{Y}\right)^2\ ;\ \text{Cov}\left(X, Y\right) = \frac{1}{n}\sum XY - \overline{X}\ \overline{Y}$$

	X	Y	X^2	Y^2	XY
	45	52	2025	2704	2340
	53	38	2805	1404	2014
	74	73	5476	5329	5402
	82	48	6724	2304	3936
	38	74	1444	5476	2812
	70	90	4900	8100	6302
Total	$\sum X = 362$	$\sum Y = 375$	$\sum X^2 = 23378$	$\sum Y^2 = 25357$	$\sum XY = 22804$

$$\overline{X} = \frac{\sum X}{n} = \frac{362}{6} = 60.3\ ;\ \overline{Y} = \frac{\sum Y}{n} = \frac{375}{6} = 62.5$$

$$\sigma_X^2 = \frac{1}{6} \times 23378 - \left(60.3\right)^2 = 260.23\ ;\ \sigma_X = 16.18$$

$$\sigma_Y^2 = \frac{1}{6} \times 25357 - \left(62.5\right)^2 = 319.92$$

$$\sigma_Y = 17.88$$

$$\text{Cov}(X,Y) = \frac{1}{n}\sum XY - \overline{X}\cdot\overline{Y}$$

$$= \frac{1}{6}\times 22804 - (60.3)(62.5) = 31.9$$

$$r = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{31.9}{16.13\times 17.88} = 0.1106 \cong 0.11$$

Regression equation of X on Y is

$$X = \overline{X} + r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right)$$

$$= 60.3 + 0.11\times\frac{16.13}{17.88}(Y - 62.5)$$

$$\boxed{X = 0.14Y + 54.05}$$

Regression equation of Y on X is

$$Y = \overline{Y} + r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right)$$

$$= 62.5 + 0.11\times\frac{17.8}{16.13}(X - 60.3)$$

$$\boxed{Y = 0.12\ X + 55.25}$$

4.8 Model Questions and Exercises:

1. Derive regression lines of Y on X and X on Y.
2. State and prove the properties of regression coefficients.
3. Define coefficient of determination.
4. Define correlation ratio and obtain the limits for it.
5. Derive the angle between two regressionlines.

6. Find the regression lines for the following data:

X	146	152	158	164	170	176	182
Y	75	78	77	79	82	85	86

7. The equations of two regression lines obtained in a correlation analysis are as follows:

$$3X + 12Y = 19 \quad ; \quad 3Y + 9X = 46$$

Obtain (i) The value of correlation coefficient

(ii) Mean values of X and Y

4.9 Summary:

Concepts of regression, regression lines, Regression Coefficients, Correlation ratio, coefficient of determination are discussed. A good number of problems are solved and some exercises are given to the students to solve on their own.

4.10 Technical Terms:

Regression

Regression Lines

Regression Coefficients

Coefficient of determination

Correlation Ratio

Lesson Writer

A. Mohan Rao

Lesson 5

SAMPLING

Objective:

After studying the lesson, the students will have clear comprehension on several basic concepts of sampling theory.

Structure of the Lesson:

5.1 Introduction:

In a statistical investigation, the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to individuals belonging to a group. The complete enumeration of the group is impossible because the group may be finite or infinite. If the group of individuals is finite, complete enumeration is not possible because the units may be destroyed in the course of inspection.

An alternate method is desirable to study the population characteristics. The inspection of the units though not destructive 100% inspection is not advisable because of many factors like administrative and monetary aspects, time factor, etc., and we take the help of sampling. If the population is infinite, complete enumeration is not possible.

For the purpose of determining population characteristics, instead of enumerating the entire population, the individuals in the sample only are observed. Then the sample characteristics are utilized to approximately determine or estimate the population characteristics. Sampling is quite often used in our day-to-day practical life. This lesson explains some basic concepts of sampling theory.

5.2 Population:

A group of individuals under study is called population. Data from the individuals with respect to some measurable characteristic is known as population or statistical population.

Example 1: Data collected from the students of a region regarding their heights can be treated as a population.

Example 2: Life time of bulbs manufactured by a company can be regarded as population. Each bulb is a primary unit in the population. The life time of a bulb is a variable since the life time varies from bulb to bulb.

Example 3: All the citizens in a city classified into graduates and non-graduates consitute a population.

5.3 Parameter:

A statistical constant obtained from a population is known as parameter of the population. In example 1, the average height is a parameter. Similarly, the average life time is a parameter in example 2. The graduates percentage in the population is a parameter in example 3.

A population is charactrised by a probability distribution and the constants of the probability distribution can be treated as the parameters of the population. If a population is assumed to be normal distribution, then the population mean and standard deviation will be parameters of the population.

The number of mistakes in a page of a book is a discrete random variable and follows Poisson distribution. The average number of mistakes per page of the book can be treated as a parameter.

5.4 Sample:

The collection of data depends upon the available resources and sometimes it is not possible to obtain the complete data regarding the study of the population. In example 1, if all the measurements of heights of students is not available within the time, complete data can not be obtained. In example 2, the experimenter has to wait till all the bulbs die out to obtain the life span of all bulbs that were put to test. It is not possible to obtain complete life span of the bulbs. Therefore it may happen to take decisions on the population characteristics with the available information. It means that a part of the population may be studied to asses the population characteristics and which leads to the following definition.

Definition:

A part of the population is called sample. A sample should be a proper representative of the population. In some investigations, the size of the population is not known but the sample size is a finite quantity.

5.5 Statistic:

In practice parameter values are unknown and the estimates based on sample values are generally used. A value obtained from the sample values is approximated to the unknown value of the parameter.

A function of sample observations is known as a statistic. It has been described that sample observations are the random variables, the function of random variables is also a random variable.

Example: Arithmetic Mean, Median, standard deviation and variance of a sample containing 'n' observations.

The value of the statistic changes as the sample values change, it may have different values as different samples can be drawn from the population. These values of the statistic constitute a probability distribution and it leads to the discussion of statistical tests of significance and estimation of population parameters.

5.6 Random Sampling and Random Numbers:

In random sampling the sample units are selected at random. A random sample is one in which each unit of population has an equal chance of being included in it.

Suppose we take a sample of size n from a finite population of size N. There will be $^{N}C_{n}$ samples. A sampling technique in which each of $^{N}C_{n}$ samples has an equal chance of being selected is known as random sampling.

A random sample can be obtained by the use of some random number tables or by using lottery method. One of the random number tables are given by Tippet. Tippet's random number tables consist of 10400 four-digited numbers giving in all 10400×4 i.e. 41600 digits. These tables have proved to be fairly random in character. Any page of the table is selected at random and the numbers in any row or column or diagonal selected at random may be taken to constitute the sample. Now using the lottery method, suppose we want to select 'r' candidates out of n. We assign 1 to n, one number to each candidate and write these numbers on n slips which are made as homogeneous as possible in shape, etc. These slips are then put in a bowl and thoroughly shuffled and then 'r' slips are drawn one by one. The 'r' candidates corresponding to the numbers on the slips drawn will constitute the random sample.

5.7 Sampling Distribution:

It has been described that a statistic is a random variable and every random variable will have a probability distribution for all the values assumed by it.

Definition:

The probability distribution of a statistic is known as a sampling distribution of a statistic.

Example 1:

If a sample of size 'n' has been drawn from a normal population with parameters μ, σ then $x_1, x_2, ..., x_n$ are identically independent random variables and any function $t_n = \bar{x}$ as a statistic which follows $t_n \sim \left(\mu, \sigma/\sqrt{n}\right)$ as the parameters and the sampling distribution of t_n is a normal distribution.

If we draw a sample of size n from a given finite population of size N, then total number of possible samples is

$$^{N}C_{n} = \frac{N!}{n!(N-n)!} = K\ (say)$$

For each of these K samples we can compute some statistic $t = t(x_1, x_2, \ldots\ldots\ldots, x_n)$, in particular the arithmetic mean $\bar{x}$, the variance s^2, etc. The set of values of the statistic so obtained, one for each sample, constitute a sampling distribution of the statistic. For example, the values $t_1, t_2, \ldots\ldots\ldots\ldots, t_k$ determine the sampling distribution of the statistic t. We can also compute the various statistical constants like mean, variance, skewness, kurtosis for its distribution.

Statistic			
SampleNo	t	$\bar{x}$	s^2
1	t_1	$\bar{x}_1$	s_1^2
2	t_2	$\bar{x}_2$	s_2^2
3	t_3	$\bar{x}_3$	s_3^2
⋮	⋮	⋮	⋮
k	t_k	$\bar{x}_k$	s_k^2

The mean and variance of the sampling distribution of the statistic t are given by

$$\bar{t} = \frac{1}{k}(t_1 + t_2 + \ldots\ldots\ldots\ldots + t_k)$$

$$= \frac{1}{k}\sum_{i=1}^{k} t_i$$

and $$\operatorname{Var}(t) = \frac{1}{k}\left[\left(t_1 - \bar{t}\right)^2 + \left(t_2 - \bar{t}\right)^2 + \ldots\ldots\ldots + \left(t_k - \bar{t}\right)^2\right]$$

$$= \frac{1}{k}\sum_{i=1}^{k}\left(t_i - \bar{t}\right)^2$$

5.8 Standard Error:

The standard deviation of the sampling distribution of a statistic is known as standard error abbreviated as S.E. The standard error is defined always for a sample statistic. S.E plays an important role in the large sample theory and forms the basis for testing of hypothesis. If t is any statistic, then for large samples,

$$Z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{t - E(t)}{S.E.(t)} \sim N(0,1)$$ for large samples.

If the discrepancy between the observed and the expected value of a statistic is greater than Z_α times its S.E., null hypothesis is rejected at α level of significance.

(i) The magnitude of the standard error gives an index of the precision of the estimate of the parameter. The reciprocal of the standard error is taken as the measure of reliability or precision of the statistic

$$S.E.(\bar{x}) = \sigma/\sqrt{n}\ .$$

The standard error of $\bar{x}$ vary inversely as the square root of the sample size. In order to double the precision which amounts to reducing the standard error to one half, the sample size has to be increased four times.

(ii) S.E. enables us to obtain the probable limits with in which the population partameter may be expected to lie. For example, the probable limits for population proportion P are given by

$$P \pm 3\sqrt{PQ/n}$$

The standard errors of some well known statistics, for large samples are given below, where n is the sample size, σ^2 the population variance, and P the population proportion. Q = 1-P, n_1 and n_2 represent the sizes of two independent random samples respectively drawn from the given population(s).

S.No.	Statistic	Standard Error
1	Sample mean	$\sigma/\sqrt{n}$
2	Observed sample proportion 'p'	$\sqrt{PQ/n}$
3	Sample S.D. : s	$\sqrt{\sigma^2/2n}$
4	Sample variance : s^2	$\sigma^2\sqrt{2/n}$
5	Difference of two sample means: $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
6	Difference of two sample S.D's $(s_1 - s_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}$
7	Difference of two sample propotions $(p_1 - p_2)$	$\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$

5.10 Examples:

Example 1: Obtain standard error of sample proportion p.

Solution: In a sample of size n, let X be the number of persons possessing the given attribute. Then

observed proportion of success = $\frac{X}{n} = p\ (\text{say})$

$$E(p) = E\left(\frac{X}{n}\right)$$

$$= \frac{1}{n}E(X).$$

The number of persons possessing an attribute (X) follows binominal distribution with mean $= nP = E(X)$.

$$\therefore\ E(p) = \frac{1}{n}\cdot nP = P\ .$$

Thus the sample proportion 'p' gives an unbiased estimate of the population proportion P.

Also $V(P) = V\left(\frac{X}{n}\right)$

$$= \frac{1}{n^2} \cdot V(X)$$

$$= \frac{1}{n^2} \cdot nPQ$$

$$V(P) = \frac{PQ}{n}$$

Standard Error (p), S.E.(p) $= \sqrt{\frac{PQ}{n}}$.

Example 2: Obtain the standard error of difference of sample proportions.

Solution: Suppose we want to compare two distinct populations with respect to the prevalence of a certain attribute A, among their members. Let X_1, X_2 be the number of persons possessing the given attribute A in random samples of sizes n_1 and n_2 from the two populations respectively. The sample proportions are defined by

$$p_1 = \frac{X_1}{n_1} \text{ and } p_2 = \frac{X_2}{n_2} \cdot$$

If P_1 and P_2 are population proportions, then

$$E(p_1) = P_1 \text{ , } E(p_2) = P_2$$

and $V(p_1) = \frac{P_1Q_1}{n_1}$ and $V(p_2) = \frac{P_2Q_2}{n_2}$.

Since for large samples, p_1 and p_2 are independently and asymptotically normally distributed, $(p_1 - p_2)$ is also normally distributed. Then

$$V(p_1 - p_2) = V(p_1) + V(p_2).$$

$\because$ p_1 and p_2 are independent sample proportions $\text{Cov}(p_1 , p_2) = 0$.

$$V(p_1 - p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}.$$

$$\therefore \text{S.E.}\left(p_1 - p_2\right) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

$\therefore$ The standard error of difference of two sample proportions is denoted by

$$\text{S.E.}\left(p_1 - p_2\right) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}\ .$$

Example 3: The standard error of the sample mean of sample size n from a population with variance σ^2 is $\sigma/\sqrt{n}$.

Solution: Let $x_1, x_2, \ldots\ldots\ldots., x_n$ be a random sample of size n from a population with variance σ^2, then the sample mean $\bar{x}$ is

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots\ldots.. + x_n\right)$$

$$V\left(\bar{x}\right) = V\left\{\frac{1}{n}\left(x_1 + x_2 + \ldots\ldots.. + x_n\right)\right\}$$

$$= \frac{1}{n^2}\left\{V\left(x_1\right) + V\left(x_2\right) + \ldots\ldots\ldots\ldots. + V\left(x_n\right)\right\}$$

$\text{Cov}\left(x_i, x_j\right) = 0, \forall i, j$, since the sample observations are independent.

$$V\left(x_i\right) = \sigma^2, \left(i = 1, 2, \ldots\ldots\ldots\ldots., n\right)$$

$$V\left(\bar{x}\right) = \frac{1}{n^2}\left\{\sigma^2 + \sigma^2 + \ldots\ldots\ldots. + \sigma^2\right\}$$

$$= \frac{1}{n^2} \cdot n\sigma^2$$

$$V\left(\bar{x}\right) = \frac{\sigma^2}{n}$$

$\therefore$ Standard error, $\text{S.E.}\left(\bar{x}\right) = \frac{\sigma}{\sqrt{n}}$.

Example 4: The standard error of difference of two sample means from normal populations.

Solution: Let $\bar{x}_1$ be the mean of a sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and let $\bar{x}_2$ be the mean of an independent random sample of size n_2 from another population with mean μ_2 and variance σ_2^2. In large samples,

$$\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \bar{x}_2 \sim N\left(\mu_2, \sigma_2^2/n_2\right)$$

$\bar{x}_1, \bar{x}_2$ being the difference of two independent normal variates is also a normal variate. $V\left(\bar{x}_1\right) = \frac{\sigma_1^2}{n_1}, V\left(\bar{x}_2\right) = \frac{\sigma_2^2}{n_2}$

$$V\left(\bar{x}_1 - \bar{x}_2\right) = V\left(\bar{x}_1\right) + V\left(\bar{x}_2\right)$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

since the sample means are independent, $\text{Cov}\left(\bar{x}_1, \bar{x}_2\right) = 0 \cdot$

$$\therefore \text{ S.E.}\left(\bar{x}_1 - \bar{x}_2\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The standard error of difference of two sample means is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Example 5: The standard error of difference of standard deviations in normal populations.

Solution: If s_1 and s_2 are the standard deviations of two independent samples, then the difference of standard deviations also a normal variate. $(s_1 - s_2) \sim$ Normal distribution

$$\text{S.E.}(s_1 - s_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}},$$

where n_1 and n_2 are the sample sizes drawn from two normal populations with variances σ_1^2 and σ_2^2 respectively.

Example 6: Derivation of sampling distribution of the statistic s^2.

Solution: In order to derive the sampling distribution of the statistic s^2, it is required to derive $E\left(s^2\right)$ and $V\left(s^2\right)$.

Let $x_1, x_2, ..., x_N$ be a population with mean $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$ and variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$.

Let $x_1, x_2, ..., x_n$. $(n < N)$ be a random sample drawn from the above population with mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Here each $x_i\ (i = 1, 2, ..., n)$ is being included in the sample at random from the population with replacement. Then each $x_i\ (i = 1, 2, ..., n)$ is a random variable which assumes one of the population values x_i with equal probabilities $P(x_i = X_i) = \frac{1}{N}, \quad (i = 1, 2, ..., n)\cdot$

Now $$E(x_i) = \sum_{i=1}^{N} X_i P(x_i = X_i)$$

$$= \frac{\sum_{i=1}^{N} X_i}{N}$$

$$= \mu \ .... (1)$$

Also $$V(x_i) = E\left[x_i - E(x_i)\right]^2$$

$$= E(x_i - \mu)^2$$

$$= \sum_{i=1}^{N} \left(x_i - \mu\right)^2 \cdot P\left(x_i = X_i\right)$$

$$= \frac{\sum_{i=1}^{N} \left(X_i - \mu\right)^2}{N}$$

$$= \sigma^2 \text{ .... } (2)$$

To derive $E\left(s^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)$

$$= \frac{1}{n}\sum_{i=1}^{n} E\left(x_i^2\right) - E\left(\bar{x}^2\right) \text{................} (3)$$

From (2) $V\left(x_i\right) = \sigma^2$

$$E\left(x_i^2\right) - \left[E\left(x_i\right)\right]^2 = \sigma^2$$

$$E\left(x_i^2\right) - \mu^2 = \sigma^2$$

$$E\left(x_i^2\right) = \sigma^2 + \mu^2 \cdots\cdots\cdots\cdots (4)$$

Also from the sampling distribution of $\bar{x}$, we have

$$V\left(\bar{x}\right) = \frac{\sigma^2}{n}$$

$$E\left(\bar{x}^2\right) - \left[E\left(\bar{x}\right)\right]^2 = \frac{\sigma^2}{n}$$

$$E\left(\bar{x}^2\right) - \mu^2 = \frac{\sigma^2}{n}$$

$$E\left(\bar{x}^2\right) = \frac{\sigma^2}{n} + \mu^2 \cdots\cdots\cdots\cdots (5)$$

Using (4) and (5) in (3), We obtain

$$E\left(s^2\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right) - \left(\frac{\sigma^2}{n}+\mu^2\right)$$

$$= \frac{n\left(\sigma^2+\mu^2\right)}{n} - \left(\frac{\sigma^2}{n}+\mu^2\right)$$

$$= \left(1-\frac{1}{n}\right)\sigma^2 \cdots\cdots\cdots\cdots\cdots(6)$$

To derive $V\left(s^2\right)$

We know that the variance of chisquare distribution with $(n-1)$ degrees of freedom is $2(n-1)$.

i.e. $V\left(\chi^2\right) = 2(n-1)$

$$V\left(\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sigma^2}\right) = 2(n-1)$$

$$V\left(\frac{ns^2}{\sigma^2}\right) = 2(n-1)$$

$$\frac{n^2}{\sigma^4}V\left(s^2\right) = 2(n-1)$$

$$V\left(s^2\right) = 2(n-1)\frac{\sigma^4}{n^2}$$

$$= \frac{2(n-1)}{n}\frac{\sigma^4}{n}$$

$$= \left(1-\frac{1}{n}\right)\frac{2}{n}\sigma^4 ....(7)$$

As $n \to \infty, \frac{1}{n}$ tends to 0 more rapidly than $\frac{2}{n}$.

Hence, the sampling distribution of s^2 in large samples has mean $E\left(s^2\right) = \sigma^2$, variance $V\left(s^2\right) = \frac{2}{n}\sigma^4$ and the standard error $S.E.\left(s^2\right) = \sqrt{V\left(s^2\right)}$.

$$= \sqrt{\frac{2}{n}\sigma^4}$$

$$= \sigma^2 \sqrt{\frac{2}{n}}$$

5.9 Sampling Distribution of Sum of Observations from Binomial Distribution:

Suppose $x_1, x_2, \ldots\ldots\ldots\ldots\ldots, x_n$ is a sample drawn from the Bernoulli population $\phi\left(x_i\right) = p^{x_i} q^{1-x_i}$, $x_i = 0,1$.

In every trial there is either a success or a failure. If there is success $x_i = 1$ and the probability of success $\phi\left(x_i\right) = p$.

If there is failure $x_i = 0$ and the probability of failure $\phi\left(x_i\right) = q = 1 - p$.

$\therefore$ The joint probability function of all values $x_1, x_2, \ldots\ldots\ldots\ldots\ldots, x_n$

= probability that the point $\left(x_1, x_2, \ldots\ldots\ldots\ldots\ldots, x_n\right)$ is obtained on the n-dimensional sample space $= \phi\left(x_1\right) \cdot \phi\left(x_2\right) \ldots\ldots\ldots\ldots \phi\left(x_n\right)$.

$$= p^{\sum x_i} q^{n - \sum x_i}$$

$$= \phi\left(x_1, x_2, \ldots\ldots\ldots, x_n\right)$$

$\because$ The sample mean $\bar{x} = \frac{1}{n}\sum x_i \cdot \left(\sum x_i \text{ varies from } 0 \text{ to } n\right)$.

$\therefore$ The possible values of $\bar{x}$ are $= 0, \frac{1}{n}, \frac{2}{n}, \ldots\ldots\ldots\ldots\ldots, \frac{k}{n}, \ldots\ldots\ldots\ldots\ldots\ldots, 1$

$\bar{x}$ takes the value $\frac{k}{n}$ when $\sum x_i = K, \left(K = 0, 1, 2, \ldots\ldots\ldots, n\right)$.

It gives $K = n\bar{x}$. For all points of the sample space having $\sum x_i = K = n\bar{x}$ the probability function is

$$\phi\left(x_1, x_2, \ldots\ldots\ldots\ldots, x_n\right) = p^k q^{n-k}.$$

We can get $\sum_{i=1}^{n} x_i = K$ in ${}^nC_K$ ways, as any K of the n values may have this values 1, and the remaining $(n-K)$ will have '0' values.

This probability function $\phi(K) = {}^nC_K\ p^K q^{n-K},\ K = 0,1,2,\ldots\ldots\ldots,n$

Sampling Distribution of Sum of Observations from Poisson distribution:

Suppose a simple random sample $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ is drawn from the Poisson distribution.

$$\phi(x) = \frac{e^{-m}\, m^x}{x!}\ (x = 0,1,2,\ldots\ldots\ldots\ldots,\infty)$$

$\therefore$ The M.G.F. of the Poisson variate $M_x\ (t) = e^{-m\left(e^t-1\right)}$

The M.G.F. of each of the variate values $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ is also same.

$\therefore$ The M.G.F. of $\left(x_1 + x_2 + \cdots\cdots\cdots\cdots + x_n\right)$ = product of all M.G.F.'s of x_i

$$= e^{-m\left(e^t-1\right)} \cdot e^{-m\left(e^t-1\right)} \cdots\cdots\cdots\cdots\cdots\, e^{-m\left(e^t-1\right)}$$

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

Which is the M.G.F. of Poisson variate having mean nm.

$\therefore$ The Poisson distribution for $x_1, x_2, \ldots\ldots\ldots\ldots, x_n = \sum x_i = n\bar{x}$

= Poisson distribution with parameter nm

$$\phi\left(n\bar{x}\right) = \frac{e^{-nm}\,(nm)^{\sum x_i}}{\left(\sum x_i\right)!}$$

where $n\bar{x} = 0,1,2,\ldots\ldots\ldots\ldots$

Sampling distribution of Sum of observations from normal population

Suppose $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ are the values of a random sample drawn from a normal population having mean μ and variance σ^2.

$\because$ $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ are independent normal variates,

$$M_{x_i(t)} = e^{mt+\frac{1}{2}\sigma^2 t^2}, \quad \text{for } i = 1, 2, \cdots\cdots\cdots, n$$

$$M_{(x_1+x_2+\cdots\cdots\cdots+x_n)}(t) = M_{x_1}(t) \cdot M_{x_2}(t) \cdots\cdots\cdots\cdots M_{x_n}(t)$$

$$= \left[e^{mt + \frac{1}{2}\sigma^2 t^2} \right]^n$$

$$= e^{nmt + \frac{1}{2}n\sigma^2 t^2}$$

$$M_{\bar{x}}(t) = M_{(x_1+x_2+\cdots\cdots+x_n)/n}(t)$$

$$= e^{nm(t/n)+\frac{1}{2}n\sigma^2(t/n)^2}$$

$$= e^{mt} \cdot + \frac{1}{2}\left(\sigma^2/_n\right) t^2$$

which is the M.G.F. of a normal variate having mean m and variance $\sigma^2/_n$.

$\therefore$ The probability differential for $\bar{x}$, when the parent population is normal,

$$dP = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-n/_2\left(\frac{\bar{x}-\mu}{\sigma}\right)^2} d\bar{x} \cdot$$

5.11 Model Questions:

1. Define the following

   (i) Population (ii) Parameter (iii) Sample (iv) Statistic

   (v) Sampling distribution (vi) Standard Error

2. Obtain the standard error of the sample mean $\bar{x}$.

3. Obtain the sampling distribution of sample mean from normal distribution.

4. Obtain the sampling distribution of sample mean from binomial distribution.

5. Find the sampling distribution of sample mean from Poisson distribution.

5.12 Summary:

The concepts of population, parameter, sample, statistic, sampling distribution and standard error are explained. The sampling distributions of sum of observations in the case of binomial, Poisson and normal distributions are derived. A good number of examples are given to obtain the standard error of various statistics. Some model questions are given to the students to prepare on their own.

5.13 Technical Terms:

Parameter

Statistic

Sampling Distribution

Standard Error

Lesson Writer

V. RamaKrishna

Lesson 6

EXACT SAMPLING DISTRIBUTIONS

Objective:

After studying the lesson the students will have clear comphrension in the theory and applications of exact sampling distributions of chi-square, t and F.

Structure of The Lesson:

6.1 Introduction:

In the definition of sampling distribution of a statistic given in Lesson 5, we discussed that every sample statistic will have a sampling distribution but in usual practice may not obtain mathematical or analytical expressions for the probability functions. But for some statistics we can find the sampling distributions and those sampling distributions are called exact sampling distributions. For the statistics sample mean, variance we have the distributions and properties which will be discussed in the following sections of this lesson.

Exact Sampling Distribution:

A sampling distribution of a statistic is said to be an exact sampling distribution, if its population follows normal distribution. Some of the important exact sampling distributions are χ^2 - distribution, Students t - distribution and Snedecor's F - distribution.

Degrees of Freedom (d.f.):

The number of independent observations, from which a statistic is computed is known as the d.f. for the statistic. The net d.f. is denoted by n - r - k where n is the number of independent sample observations, 'r' is the number of restrictions imposed on the statistic and k is the number of unknown parameters of the population.

6.2 Chi-square Distribution:

If $X_i \left(i=1,2,............,n\right)$ are n independent normal variates with mean μ_i and variances $\sigma_i^2 \left(i=1,2,............,n\right)$, then

$$\chi_n^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2$$ is a chi-square variate with 'n' degrees of freedom.

If a random variable X has a chi-square distribution with n d.f., we write $X \sim \chi^2_{(n)}$ and its p.d.f. is

$$f(x) = \frac{1}{2^{n/2}\left|\overline{(n/2)}\right.} e^{-x/2} x^{\frac{n}{2}-1}, 0 \le x < \infty$$

M.G.F. of Chi-square distribution is

$$M_X(t) = (1-2t)^{-n/2}, |2t| < 1$$

Cumulant Generating function

$$K_X(t) = \log M_X(t)$$

$$= \frac{-n}{2}\log(1-2t)$$

$$= \frac{-n}{2}\left[2t + \frac{(2t)^2}{2!} + \frac{(2t)^3}{3!} + ...\right]$$

$$K_1 = \text{Coeff of } t \text{ in } K(t) = n$$

$$K_2 = \text{Coeff of } \frac{t^2}{2!} \text{ in } K(t) = 2n$$

$$K_3 = \text{Coeff of } \frac{t^3}{3!} \text{ in } K(t) = 8n$$

$$K_4 = \text{Coeff of } \frac{t^4}{4!} \text{ in } K(t) = 48n$$

$$\text{Mean} = K_1 = n$$

$$\text{Variance} = K_2 = 2n$$

$$\mu_3 = K_3 = 8n$$

$$\mu_4 = K_4 + 3K_2^2 = 48n + 12n^2$$

$$\beta_1 = \frac{8}{n}$$

$$\beta_2 = \frac{12}{n} + 3.$$

Characteristic function $\phi_x(t) = (1-2it)^{-n/2}$

Mode of χ^2 - distribution = n - 2

skewness $= \sqrt{\frac{2}{n}}$

When $n = 1, \chi^2$ distribution become standard normal distribution. χ^2 - distribution is asymptotically normally distributed. i.e., $\chi^2 \rightarrow$ Normal as $n \rightarrow \infty$.

Additive Property:

The sum of independent χ^2 - variates is also a χ^2 - variate. More precisely,

If $\chi_i^2 \; (i = 1, 2, \cdots\cdots\cdots\cdots, k)$ are independent χ^2 - variates with n_i degrees of freedom respectively, then the sum $\sum_{i=1}^{k} \chi_i^2$ is also a χ^2 - variate with $\sum_{i=1}^{k} n_i$ degrees of freedom.

i.e., $M_{\sum_{i=1}^{k} \chi_i^2}(t) = (1-2t)^{-(n_1+n_2+\cdots\cdots+n_k)/2}$.

Applications of Chi-square distribution:

χ^2 - distribution has a large number of applications in statistics, some of them are given below:

i) To test if the hypothetical value of the population variance $\sigma^2 = \sigma_0^2$

ii) To test the goodness of fit between observed and expected frequencies.

iii) To test the independence of the attributes.

iv) To test the homogeneity of independent estimates of the population variance.

v) To test the homogeneity of independent estimates of the population correlation coefficients.

vi) To combine various probabilities obtained from independent experiments to give a single test of significance.

6.3 Student's t-distribution:

Let $x_i\left(i=1,2,...........,n\right)$ be a random sample of size n from a normal population with mean μ and variance σ^2. Then Student's t is defined by a statistic

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

where $\overline{x}=\frac{1}{n}\sum x_i$ is the sample mean and $S^2=\frac{1}{n-1}\sum_i\left(x_i-\overline{x}\right)^2$ is an unbiased estimate of the population variance σ^2, and it follows Student's t distribution with $\nu=n-1$ degrees of freedom with probability density function

$$f\left(t\right)=\frac{1}{\sqrt{\nu}\,\beta\left(\frac{1}{2},\frac{\nu}{2}\right)}\frac{1}{\left(1+\frac{t^2}{\nu}\right)^{(\nu+1)/2}},\ -\infty<t<\infty\ .$$

Fisher's - t:

The ratio of a standard normal variate to the square root of an independent chi-square variate divided by its degrees of freedom is defined as Fisher's - t statistic. If □ is a N (0,1) and χ^2 is an independent chi-square variate with n d.f, then Fisher's t is given by

$$t=\frac{\square}{\sqrt{\chi^2/n}}$$ and it follows Student's t distribution with n degrees of freedom.

Constants of t - distribution:

$f\left(t\right)$ is symmetrical about the line t = 0, all the moments of odd order about origin vanish.

$$\mu_{2r+1}^1\left(\text{about origin}\right)=0,\ r=0,1,2,..............$$

μ_1^1 (about origin) = 0 = Mean

The central moments coincide with moments about origin.

$$\mu_{2r+1} = 0 \ (r = 1, 2, ...)$$

The moments of even order are given by

$$\mu_2 = \frac{n}{n-2} \cdot , (n > 2)$$

$$\mu_4 = \frac{3n^2}{(n-2)(n-4)}, \ (n > 4)$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3\left(\frac{n-2}{n-4}\right), \ (n > 4)$$

t - distribution tends to normal distribution when $\nu \to \infty$.

Student's t regarded as a particular case of Fisher's t:

Since $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$

$$\square \ = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

and $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{\sigma^2}$

is independently distributed as chi - square variate with $(n-1)$ d - f.

Hence Fisher's t is given by

$$t = \frac{\square}{\sqrt{\chi^2/(n-1)}} = \frac{\sqrt{n}\,(\bar{x} - \mu)}{\sigma} = \frac{\sigma}{\sqrt{\frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2}}$$

$$= \frac{\sqrt{n}\,(\bar{x} - \mu)}{S}$$

$$= \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

and it follows Student's t - distribution with $(n-1)$ degrees of freedom.

Applications of t - distribution:

The t - distribution has a wide number of applications in statistics, some of them are given below:

i) t - test is used to test if the sample mean $\left(\bar{x}\right)$ differs significantly from the hypothetical value μ of the population mean.

ii) t - test is used to test the significance of the difference between two sample means.

iii) t - test is used to test the significance of an observed sample correlation coefficient and sample regression coefficient.

iv) t - test is used to test the significance of observed partial correlation coefficient.

6.4 F - Distribution:

If X_1 and X_2 are two independent chi-square variates with n_1 and n_2 degrees of freedom respectively, then F - statistic is defined by

$$F=\frac{X_1/n_1}{X_2/n_2}.$$

F is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and it follows Snedecor's F - distribution with $\left(n_1, n_2\right)$ degrees of freedom with probability function given by

$$f\left(F\right)=\frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{\beta\left(\frac{n_1}{2},\frac{n_2}{2}\right)}\;\frac{F^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1}{n_2}F\right)^{\frac{\left(n_1+n_2\right)}{2}}},\;0\le F<\infty .$$

Constants of F - distribution:

$$\mu_r^1=\left(\frac{n_2}{n_1}\right)^r\;\frac{\Gamma\left[r+\left(\frac{n_1}{2}\right)\right]\Gamma\left[\left(\frac{n_2}{2}\right)-r\right]}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}$$

In particular $\mu_1^1=\frac{n_2}{n_2-2},\; n_2>2$

$$\mu_2^1 = \frac{n_2^2(n_1+2)}{n_1(n_2-2)(n_2-4)},\ n_2 > 4$$

$$\mu_2 = \frac{2n_2^2(n_2+n_1-2)}{n_1(n_2-2)^2(n_2-4)},\ n_2 > 4\cdot$$

Mode of F distribution $= \frac{n_2(n_1-2)}{n_1(n_2+2)}$.

The points of inflexion of F - distribution exist for $n_1 > 4$ and are equidistant from mode.

Karl pearson's coefficient of skewness $S_K = \frac{\text{Mean} - \text{Mode}}{\sigma} > 0$

since mean > 1 and mode < 1, F distribution is highly positively skewed.

The probability f(F) increases steadily at first until it reaches its peak and then decreases slowly so as to become tangential at $F = \infty$. F - axis is an asymtote to the right tail.

Applications of F - Distribution:

1. F - test is used to test the equality of two population variances.
2. F - test is used to test the significance of an observed Multiple Correlation Coefficient.
3. F - test is used to test the significance of an observed sample correlation ratio.
4. F - test is used to test the linearity of a fitted regression line.
5. F - test is used to test the equality of several means.

6.5 Relation between t and F Distributions:

In F - distribution with (n_1, n_2) degrees of freedom take $n_1 = 1, n_2 = n$ and $t^2 = F$,

$$dF = 2tdt$$

The probability differential of F transforms to

$$dG(t) = \frac{\left(\frac{1}{n}\right)^{\frac{1}{2}}}{\beta\left(\frac{1}{2}, \frac{n}{2}\right)} \frac{\left(t^2\right)^{\frac{1}{2}-1}}{\left(1+\frac{t^2}{n}\right)^{\frac{(n+1)}{2}}} 2tdt,\quad 0 \le t^2 < \infty$$

$$= \frac{1}{\sqrt{n}\,\beta\left(\frac{1}{2}, \frac{n}{2}\right)} \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}} \text{dt}, \quad -\infty < t < \infty \cdot$$

This is the probability function of Students t - distribution with n degrees of freedom. We have the following relation between t and F distributions.

If a statistic t follows t - distribution with n degrees of freedom, then t^2 follows Snedecor's F - distribution with $(1,n)$ degrees of freedom. Symbolically

If $t \sim t(n)$

then $t^2 \sim F(1,n)$.

6.6 Relation between F and χ^2 Distributions:

In $F(n_1, n_2)$ distribution if we let $n_2 \to \infty$ then, $\chi^2 = n_1 F$ follows χ^2 - distribution with n_1 degrees of freedom.

$$f(F) = \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} F^{\frac{n_1}{2}-1}}{\Gamma\left(n_1/2\right)\Gamma\left(n_2/2\right)} \frac{\Gamma\left[(n_1+n_2)/2\right]}{\left(1+\frac{n_1}{n_2}F\right)^{\frac{(n_1+n_2)}{2}}}, \; 0 < F < \infty$$

If $n_2 \to \infty$, we have

$$\frac{\Gamma\left[(n_1+n_2)/2\right]}{n_2^{n_1/2}\,\Gamma\left(n_2/2\right)} \to \frac{\left(n_2/2\right)^{\frac{n_1}{2}}}{n_2^{n_1/2}} = \frac{1}{2^{n_1/2}} \text{.......................}(1)$$

$$\left[\because \; \frac{\Gamma(n+K)}{\Gamma(n)} \to n^K \quad \text{as} \quad n \to \infty\right]$$

Also $\lim\limits_{n_2\to\infty}\left(1+\frac{n_1}{n_2}F\right)^{\frac{(n_1+n_2)}{2}} = \lim\limits_{n_2\to\infty}\left[\left(1+\frac{n_1}{n_2}F\right)^{n_2}\right]^{\frac{1}{2}} \times \lim\limits_{n_2\to\infty}\left(1+\frac{n_1}{n_2}F\right)^{\frac{n_1}{2}}$ ....(2)

On using (1) and (2), the p.d.f. of $\chi^2 = n_1F$ becomes

$$dP\left(\chi^2\right) = \frac{\left(n_1/2\right)^{n_1/2} e^{-\chi^2/2}}{\Gamma\left(n_1/2\right)} \left(\frac{\chi^2}{n_1}\right)^{\frac{n_1}{2}-1} d\left(\frac{\chi^2}{n_1}\right)$$

$$= \frac{1}{2^{n_1/2}\,\Gamma\left(n_1/2\right)} e^{\chi^2/2} \left(\chi^2\right)^{\frac{n_1}{2}-1} d\chi^2, \quad 0 < \chi^2 < \infty,$$

which is the p.d.f. of chi-square distribution with n_1 degrees of freedom.

6.7 Model Questions:

1. Explain F - distribution. Derive the relationships of t & χ^2 with F - distribution.
2. Explain t - distribution and write the properties and applications.
3. Explain F - distribution and write the applications and properties.
4. Define exact sampling distributions and obtain the relationships of t, χ^2 with F - distribution.
5. Define χ^2 - variate, χ^2 - distribution and mention the constants of χ^2 distribution and applications.

6.8 Summary:

The properties and applications of exact sampling distributions t, χ^2 and F distributions are covered in the lesson. Also the relationship between t and F distributions, the relationship between F and χ^2 distributions are discussed. Some model questions are given to the students to prepare on their own.

6.9 Technical Terms:

Exact sampling distribution

χ^2 statistic

Students - t statistic

F - statistic.

Lesson Writer

V. Ramakrishna

Lesson 7

THEORY OF ESTIMATION

Objective:

After studying the lesson the students will be conversant with the basic concepts of theory of estimation in statistical inference, point estimation, criteria of good estimatior.

Structure of the Lesson:

7.1 Introduction:

The theory of estimation was introduced by Prof. R.A. Fisher around 1930. The concept deals with estimation of parameters in a population. The parameters are estimated by different estimating functions known as estimators. Here we study about a good estimator and its characteristics because a parameter can be estimated by different estimators.

7.2 Problem of Estimation:

Suppose a population has a probability function with certain parameters. Knowing about unknown parameters with the known sample values is called the problem of estimation.

Example: If $f(x, \theta)$ is the probability density function of a population with θ as an unknown parameter. Then knowing about θ with the sample values is a problem of estimation.

7.3 Point Estimation:

If a parameter is represented by a point which is a value of a function of sample values, then it is called a problem of point estimation. This function of sample values is known as an estimator for the unknown parameter. A particular value of the estimator is referred as an estimate.

To estimate the unknown parameter of the population, we can use several estimators (statistics). But a statistic that determines the true value of the parameter is known as a good estimator. In other words the statistic whose distribution concentrates as closely as possible near the true value of the parameter may be the best estimate. Here we have to study the criteria of a good estimator since there exists different estimators to estimate the population parameter.

7.4 Criteria of a good estimator:

Suppose we have a sample $x_1, x_2, ..., x_n$ of size 'n' from a population with probability function $f(x; \theta_1, \theta_2, ..., \theta_K)$ where $\theta_1, \theta_2, ..., \theta_K$ are the unknown parameters of the population. If $t_n = t(x_1, x_2, ..., x_n)$ is a statistic to estimate a parameter θ then t_n is said to be a good estimator if it possess the following properties:

1) Consistency
2) Unbiasedness
3) Efficiency
4) Sufficiency

7.5 Consistency:

An estimator $t_n = t(x_1, x_2, ..., x_n)$ based on a random sample of size n is said to be a consistent estimator of parameter θ, if t_n converges to θ in probability.

t_n is a consistent estimator of θ if for every $\in > 0,$ there exists a positive integer n such that

$$P\{|t_n - \theta| < \in\} \to 1 \text{ as } n \to \infty .$$

Conditions for Consistency:

If $\{t_n\}$ be a sequence of estimators such that

i) $E(t_n) \to \theta$ as $n \to \infty$,

and ii) $V(t_n) \to 0$ as $n \to \infty$.

Example 1: If a sample of size n drawn from a normal distribution with parameters μ and σ^2, then the sample mean $\bar{x}$ is a consistent estimator of μ.

Solution:

If $x_1, x_2, ..., x_n$ is a random sample then sample arithmetic mean is $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$.

$x_1, x_2, ..., x_n$ are i.i.d. random variables, then

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Let $Z = \bar{x} - \mu$, then $Z \sim N\left(0, \frac{\sigma^2}{n}\right)$

$$\lim_{n \to \infty} P\left\{\left|\bar{x} - \mu\right| > \in\right\} = \lim_{n \to \infty} P\left\{\left|Z\right| > \in\right\}$$

$$= 1 - \lim_{n \to \infty} P\left\{\left|Z\right| \le \in\right\}$$

$$= 1 - \lim_{n \to \infty} \int_{-\in}^{\in} \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}Z^2} dZ$$

$$= 1 - \lim_{n \to \infty} \int_{-\in\sqrt{n}/\sigma}^{\in\sqrt{n}/\sigma} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}y^2} dy$$

$$= 1 - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \quad \text{where} \quad y = \frac{Z\sqrt{n}}{\sigma}, \quad dy = dZ\frac{\sqrt{n}}{\sigma}$$

$$= 1 - 1 = 0.$$

$$\Rightarrow \lim_{n \to \infty} P\left\{\left|\bar{x} - \mu\right| \le \in\right\} = 1 \cdot$$

$\therefore$ $\bar{x}$ is a consistent estimator of μ.

Example 2: The sample variance s^2 is a consistent estimator of the population variance σ^2 in the case of sample drawn from a normal population.

Solution: The sample variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

The sampling distribution of $\frac{ns^2}{\sigma^2}$ is a χ^2 with $(n-1)$ degrees of freedom.

$$E\left(\frac{ns^2}{\sigma^2}\right) = n-1$$

$$V\left(\frac{ns^2}{\sigma^2}\right) = 2(n-1)$$

$$E\left(s^2\right) = \left(\frac{n-1}{n}\right)\sigma^2$$

$$V\left(s^2\right) = 2\left(\frac{n-1}{n^2}\right)\sigma^4.$$

$$\lim_{n\to\infty} E\left(s^2\right) = \lim_{n\to\infty}\left(1-\frac{1}{n}\right)\sigma^2 = \sigma^2$$

$$\lim_{n\to\infty} V\left(s^2\right) = \lim_{n\to\infty}\left(\frac{1}{n}-\frac{1}{n^2}\right)2\sigma^4$$

$$= 0.$$

$\therefore$ From the conditions of consistency, s^2 is a consistent estimator of σ^2.

7.6 Unbiasedness:

If average value of an estimator is equal to the true value of the population parameter then the estimator is an unbiased estimator and the property of estimator is unbiasedness.

Unbiased Estimator:

An estimator $t_n = t(x_1, x_2, \ldots\ldots\ldots, x_n)$ is said to be an unbiased estimator of parameter θ, if

$$E(t_n) = \theta.$$

Example 1: In sampling from a normal population with mean μ and variance σ^2, if

$$E\left(\bar{x}\right) = \mu,$$

then the sample mean is unbiased estimator of population mean μ.

Example 2: Let $x_1, x_2, ..., x_n$ be a random sample from a normal population $N(\mu,1)$. Show that $t = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ is an unbaised estimator of $\mu^2 + 1$.

Solution: We are given $E(x_i) = \mu, \forall\ i = 1, 2, ...n$

$$V(x_i) = 1$$

$$V(x_i) = E\left(x_i^2\right) - \{E(x_i)\}^2$$

$$E\left(x_i^2\right) = V(x_i) + \{E(x_i)\}^2$$

$$= 1 + \mu^2$$

$$\therefore E(t) = E\left(\frac{1}{n}\sum x_i^2\right)$$

$$= \frac{1}{n}\sum E\left(x_i^2\right)$$

$$= \frac{1}{n}\cdot\sum_{i=1}^{n}\left(1 + \mu^2\right)$$

$$= 1 + \mu^2$$

Hence t is an unbiased estimatior of $1 + \mu^2$.

7.7 Bias and Mean Square Error:

Bias of estimator:

If $E(T_n) \neq \theta$, then T_n is called a biased estimator of the unknown parameter θ. The amount of bias $b = E(T_n) - \theta$.

Bias will be positive if $E(T_n) > \theta$, and negative

if $E(T_n) < \theta$.

Mean - Square Error (M.S.E.):

If there are more than one unbiased estimator, the problem arises which one to choose out of the class of unbiased estimators. Not only this, one aspires that sampling variance as well as bias should be minimum. These problems are tackled with the help of mean - squared error (M.S.E.). The mean squared error of an estimator T_n of θ is given as

$$M \cdot S.E. = E(T_n - \theta)^2$$

$$= E\left[E(T_n) + T_n - \theta - E(T_n)\right]^2$$

$$= E\left[\{E(T_n) - \theta\} + \{T_n - E(T_n)\}\right]^2$$

$$= E\left[E(T_n) - \theta\right]^2 + E\left[T_n - E(T_n)\right]^2$$

$$= (\text{bias})^2 + V(T_n)$$

where $E(T_n) - \theta$ is bias.

Mean square error will be minimum if T_n is an unbiased estimator of θ. i.e., $E(T_n) = \theta$ and $\text{Var}(T_n)$ is minimum. It is a difficult task to have an estimator with least mean square error. Hence, it is our endeavour to search for an estimator with uniformly minimum variance among the class of unbiased estimatiors.

7.8 Minimum Variance Unbiased (M.V.U.) Estimator:

If a statistic $t = t(x_1, x_2, \ldots\ldots\ldots, x_n)$, based on a sample of size n is such that

i) t is unbiased for θ, and

ii) It has the smallest variance among the class of all unbiased estimators of θ, then t is called the minimum variance unbiased estimator (MVUE) of θ.

More precisely, t is MVUE of θ if

$$E(t) = \theta,$$

and $$\text{Var}(t) \le \text{Var}\left(t'\right),$$

where t' is any other unbiased estimator of θ.

7.9 Efficiency:

In general, there may exist more than one consistent estimator of a parameter and also in the case of unbiasedness there may exist more than one unbiased estimator to the population parameter. There is a necessity of some further criterion which will enable us to choose between the estimators with the common property of consistency. The criterion is efficiency based on the variances of the sampling distribution of estimators.

If there exists two consistent estimators T_1 and T_2 to estimate the unknown parameter θ.

and $V(T_1) < V(T_2)$ for all n,

then T_1 is more efficient than T_2 for all sample sizes.

Most Efficient Estimator:

If in a class of consistent estimators for a parameter, there exists one estimator whose sampling variance is less than that of any other estimator then it is called the most efficient estimator.

If T_1 is the most efficient estimator with variance V_1 and T_2 is any other estimator with variance V_2 then the efficiency, E of T_2 is defined as

$$E = \frac{V_1}{V_2}$$

E can not exceed unity.

7.10 Sufficiency and Neyman Factorization Theorem:

Likelihood Function:

Let $x_1, x_2, ..., x_n$ be a random sample of size n from a population with density funtion $f(x, \theta)$. Then the likehood function of the sample observations $x_1, x_2, ..., x_n$ usually denoted by L is the joint density function of sample values.

$$L = f(x_1, \theta)\ f(x_2, \theta) .... f(x_n, \theta)$$

$$= \prod_{i=1}^{n} f(x_i, \theta).$$

L gives the relative likelihood that the random variables assume a particular set of values $x_1, x_2, ..., x_n$.

An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter.

If $T = t(x_1, x_2, ..., x_n)$ is an estimator of a parameter θ, based on a sample $x_1, x_2, ..., x_n$ of size n from the population with density $f(x, \theta)$ such that the conditional distribution of $x_1, x_2, ..., x_n$ given T, is independent of θ, then T is sufficient estimator for θ.

Neyman introduced factorization theorem and it provides the necessary and sufficient condition for a distribution to admit sufficient statistic.

Statement:

$T = t(x)$ is sufficient for θ if and only if the joint density function L of the sample values can be expressed in the form

$$L = g_\theta\left[t(x)\right] h(x),$$

where $g_\theta\left[t(x)\right]$ depends on θ and x only through the value of $t(x)$ and $h(x)$ is independent of θ.

7.11 Derivation of Sufficient Statistics:

7.11.1 Derivation of Sufficient Statistics in Normal Distribution:

The likelihood equation of sample observations is

$$L = \prod_{i=1}^{n} f\left(x_i, \mu, \sigma^2\right)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{\frac{-1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right)\right\}$$

$$= g\left[t(x)\right]\cdot h(x)$$

where $g\left[t(x)\right] = \left(\frac{1}{\sigma}\right)^n \exp\left[\frac{-1}{2\sigma^2}\left\{t_2(x) - 2\mu t_1(x) + n\mu^2\right\}\right]$

$$t(x) = \left[t_1(x),\ t_2(x)\right] = \sum x_i,\ \sum x_i^2$$

and $h(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^n$

Thus by statement of Factorization Theorem the likelihood equation can be written as the product terms one depending on parameters and another independent of the parameters.

$\therefore\ t_1(x) = \sum x_i$ is sufficient for μ

and $t_2(x) = \sum x_i^2$ is sufficient for σ^2.

7.11.2 Derivation of Sufficient Statistic in Exponential Distribution:

Let $x_1, x_2, \ldots\ldots\ldots, x_n$ be a random sample drawn from an exponential population with density function

$$f(x) = \theta\, e^{-\theta x}.$$

The likelihood function of sample observations is

$$L(x_1, x_2, \ldots\ldots\ldots, x_n) = f(x_1),\ f(x_2) \ldots\ldots\ldots\ldots f(x_n)$$

$$= \theta e^{-\theta x_1}\ \theta e^{-\theta x_2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ \theta e^{-\theta x_n}$$

$$= \theta^n\ e^{-\theta \sum_{i=1}^{n} x_i}$$

$$= \left(\theta^n\ e^{-\theta \sum x_i}\right) \times 1$$

where $g\left(t = \sum x_i, \theta\right) = \theta^n\ e^{-\theta \sum_{i=1}^{n} x_i}$ and $h(x) = 1$

which does not involve the parameter θ.

Hence by the factorization theorem the statistic $\sum x_i$ is a sufficient estimator for the parameter θ.

7.11.3 Derivation of sufficient statistic for the Binomial Distribution Parameter P:

Let $(x_1, x_2, \ldots\ldots\ldots, x_n)$ be a random sample drawn from a Binomial Population $P(X = x_i) = {}^nC_{x_i} P^{x_i}\ (1-P)^{n-x_i}$. Consider the likelihood function,

$$L(x_1, x_2, \ldots\ldots x_{i,} \ldots\ldots, x_n) = P(x_1)\ P(x_2) \ldots\ldots\ldots\ldots\ldots P(x_n)$$

$$= {}^nC_{x_1} P^{x_1} (1-P)^{n-x_1}\ {}^nC_{x_2} P^{x_2} (1-P)^{n-x_2} \ldots\ldots\ldots\ {}^nC_{x_n} P^{x_n} (1-P)^{n-x_n}$$

$$=\left(\prod_{i=1}^{n} {}^{n}C_{x_i}\right) P^{\sum_{i=1}^{n} x_i} \left(1-P\right)^{\sum_{i=1}^{n}(n-x_i)}$$

$$=\left(P^{\sum_{i=1}^{n} x_i} \left(1-P\right)^{n^2-\sum_{i=1}^{n} x_i}\right)\left(\prod_{i=1}^{n} {}^{n}C_{x_i}\right)$$

$$=g\left(x,\, t=\sum x_i,\, \theta=P\right)=P^{\sum x_i}\left(1-P\right)^{n^2-\sum x_i}$$

$h\left(x\right)=\prod_{i=1}^{n} {}^{n}C_{x_i}$ which is independent of the parameter P.

By factorization theorem, $\sum x_i$ is a sufficient statistic to the parameter P.

7.11.4 Derivation of sufficient statistic for the Poisson distribution parameter λ:

Let $x_1, x_2, \ldots\ldots\ldots, x_n$ be a random sample drawn from Poisson population with

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

Then x_i is identically independently distributed with $P\left(X=x_i\right)=\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$.

Now $x_1+x_2+\ldots\ldots\ldots+x_n=\sum x_i=t$ is a sum of n independent Poisson variates follows Poisson distribution with parameter $\left(\lambda+\lambda+\lambda+\ldots\ldots\ldots\ldots\ldots+\lambda\right)=n\lambda$,

$$P\left(t=\sum x_i\right)=\frac{e^{-n\lambda}\left(n\lambda\right)^{\sum x_i}}{\left(\sum x_i\right)!}.$$

To show that t is sufficient estimator for λ, it is required to show that

$P\left\{x_1 \cap x_2 \cap \ldots\ldots\ldots\ldots \cap x_n | t\right\}$ is independent of the parameter λ.

$$=\frac{P\left(x_1 \cap x_2 \cap \ldots\ldots\ldots \cap x_i \cap \ldots\ldots\ldots\ldots \cap x_n\right) \cap t=\sum x_i}{P\left(t=\sum x_i\right)}$$

$\{\because x_1 \cap x_2 \cap ....\cap x_n$ is subset of $x_1 + x_2 + ... + x_n\}$

$$= \frac{P(x_1 \cap x_2 \cap ....\cap x_n)}{P(t = \sum x_i)}$$

$$= \frac{P(x_1)\, P(x_2)...............P(x_n)}{P(t = \sum x_i)}$$ $\therefore x_i$'s are independent.

$$= \frac{\frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdots\cdots\cdots\cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}}{\frac{e^{-n\lambda}(n\lambda)^{\sum x_i}}{(\sum x_i)!}} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod\limits_{i=1}^{n} x_i!} \times \frac{(\sum x_i)!}{e^{-n\lambda} n^{\sum x_i} \lambda^{\sum x_i}}$$

$$= \frac{(\sum x_i)!}{\prod\limits_{i=1}^{n} x_i! \times n^{\sum x_i}}$$ which is independent of λ.

Hence $t = \sum x_i$ is a sufficient statistic for the poission parameter λ.

7.12 Examples:

Example 1: A random sample $x_1, x_2, ..., x_5$ of size 5 is drawn from a normal population with unknown mean μ and variance σ^2. Consider the following estimators to estimate μ.

i) $t_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$ ii) $t_2 = \frac{X_1 + X_2}{2} + X_3$

iii) $t_3 = \frac{2X_1 + X_2 + \lambda X_3}{3}$

where λ is such that t_3 is an unbiased estimator of μ.

a) Find λ

b) Are t_1 and t_2 unbiased?

c) State giving reasons, the estimator which is best among t_1, t_2 and t_3?

Solution:

Given $E(x_i)=\mu,\ V(x_i)=\sigma^2$

$$Cov(X_i, X_j)=0,\ (i\neq j=1,2,............,n)$$

i) $E(t_1)=E\left(\frac{X_1+X_2+..........+X_5}{5}\right)=\frac{1}{5}\sum_{i=1}^{5}E(X_i)=\frac{1}{5}\sum_{i=1}^{n}\mu=\frac{1}{5}\cdot 5\mu=\mu$

ii) $E(t_2)=E\left(\frac{X_1+X_2}{2}+X_3\right)=E\left(\frac{X_1+X_2}{2}\right)+E(X_3)$

$$=\frac{1}{2}\left[E(X_1)+E(X_2)\right]+E(X_3)$$

$$=\frac{1}{2}[\mu+\mu]+\mu=2\mu$$

iii) $E(t_3)=\mu\cdot$ $\quad\left[\therefore t_3 \text{ is an unbiased estimator of } \mu\right]$

$$\Rightarrow E\left(\frac{2X_1+X_2+\lambda X_3}{3}\right)=\mu$$

$$\frac{1}{3}\left[2E(X_1)+E(X_2)+\lambda E(X_3)\right]=\mu$$

$$\frac{1}{3}[2\mu+\mu+\lambda\mu]=\mu$$

$$3\mu+\lambda\mu=3\mu$$

$$\lambda\mu=0$$

$$\therefore \lambda=0$$

$$V(t_1)=\frac{1}{25}\left[V(X_1)+V(X_2)+V(X_3)+V(X_4)+V(X_5)\right]$$

$$=\frac{1}{25}\cdot 5\sigma^2=\frac{1}{5}\sigma^2$$

$$V(t_2) = \frac{1}{4}\left[V(X_1) + V(X_2)\right] + V(X_3)$$

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

$$V(t_3) = \frac{1}{9}\left[4V(X_1) + V(X_2)\right] = \frac{1}{9}\left(4\sigma^2 + \sigma^2\right)$$

$$= \frac{5}{9}\sigma^2$$

t_1 is an unbiased estimator but t_2 is not unbiased estimator. t_1 is the best estimator of μ since it is unbiased and has the least variance when compared to that of other estimators t_2 and t_3.

Example 2: If T is an unbiased estimator for θ, show that T^2 is a biased estimator for θ^2.

Solution: Since T is an unbiased estimator for θ, we have

$$E(T) = \theta$$

$$V(T) = E\left(T^2\right) - \left[E(T)\right]^2$$

$$= E\left(T^2\right) - \theta^2$$

$$\Rightarrow \quad E\left(T^2\right) = \mathrm{Var}(T) + \theta^2$$

$\therefore \; E\left(T^2\right) \neq \theta^2$, T^2 is a biased estimator for θ^2.

Example 3: If $x_1, x_2, ..., x_n$ are random observations on a Bernoulli variate X taking the value 1 with probability P and the value 0 with probability (1-P), show that

$\frac{\sum x_i}{n}\left(1 - \frac{\sum x_i}{n}\right)$ is a consistent estimator of $P(1-P)$.

Solution: Since $x_1, x_2, ..., x_n$ are i.i.d. Bernoulli variates with parameter 'P'

$$T = \sum_{i=1}^{n} x_i \sim B(n, P)$$

$$E(T) = nP, \quad V(T) = nPq$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{T}{n}$$

$$E(\overline{X}) = \frac{1}{n}E(T)\cdot = \frac{1}{n}\cdot nP = P$$

and $$V(\overline{X}) = V\left(\frac{T}{n}\right) = \frac{1}{n^2}\cdot V(T) = \frac{1}{n^2}\cdot nPq = \frac{Pq}{n}$$

$$V(T) = \frac{Pq}{n} \to 0 \text{ as } n \to \infty$$

$\because$ $E(\overline{X}) \to P$ and $V(\overline{X}) \to 0$ as $n \to \infty$

$\overline{X}$ is a consistent estimator of P.

$\frac{\sum x_i}{n}\left(1 - \frac{\sum x_i}{n}\right) = \overline{X}(1-\overline{X})$ being a polynomial in $\overline{X}$, is a continuous function of $\overline{X}$.

$\because \overline{X}$ is a consistent estimator of P, by the invariance property of consistent estimators

$\overline{X}(1-\overline{X})$ is a consistent estimator of $P(1-P)$.

Example 4: If T_1 and T_2 be two unbiased estimators of θ with variances σ_1^2, σ_2^2 and correlation ρ, what is the best unbiased linear combination of T_1 and T_2 and what is the variance of such a combination?

Solution: Let T_1 and T_2 be two unbiased estimators of θ.

$$E(T_1) = E(T_2) = \theta \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Let T be a linear combination of T_1 and T_2

given by $T = \ell_1 T_1 + \ell_2 T_2$

where ℓ_1, ℓ_2 are arbitrary constants.

$$E(T) = \ell_1 E(T_1) + \ell_2 E(T_2)$$

$$= (\ell_1 + \ell_2)\theta \qquad \text{(From (1))}$$

$\therefore T$ is also an unbiased estimator of θ if and only if

$$\ell_1 + \ell_2 = 1 ....(2)$$

$$V(T) = V(\ell_1 T_1 + \ell_2 T_2)$$

$$= \ell_1^2 V(T_1) + \ell_2^2 V(T_2) + 2\ell_1\ell_2 \text{Cov}(T_1, T_2)$$

$$= \ell_1^2\sigma_1^2 + \ell_2^2\sigma_2^2 + 2\ell_1\ell_2\, \rho\, \sigma_1\sigma_2 \cdots\cdots\cdots\cdots(3)$$

We want the minimum value of (3) for variations in ℓ_1 and ℓ_2 subject to the condition (2).

$$\frac{\partial}{\partial \ell_1} V(T) = 0 = \ell_1\sigma_1^2 + \ell_2\rho\sigma_1\sigma_2$$

$$\frac{\partial}{\partial \ell_2} V(T) = 0 = \ell_2\sigma_2^2 + \ell_1\rho\, \sigma_1\sigma_2$$

Subtracting we get

$$\ell_1\left(\sigma_1^2 - \rho\sigma_1\sigma_2\right) = \ell_2\left(\sigma_2^2 - \rho\sigma_1\sigma_2\right)$$

$$\frac{\ell_1}{\sigma_2^2 - \rho\sigma_1\sigma_2} = \frac{\ell_2}{\sigma_1^2 - \rho\sigma_1\sigma_2} = \frac{\ell_1 + \ell_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} = \frac{1}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

$$\therefore \ell_1 = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \text{ and } \ell_2 = \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

with the values of ℓ_1 and ℓ_2, T is the best unbiased linear combination of T_1 and T_2 and its variance is given by (3).

7.13 Model Questions and Exercises:

1. Give the criteria of a good estimator and prove that in a random sampling from $N(\mu, \sigma^2)$ population, the sample mean is unbiased estimator of μ.

2. Suppose $x_1, x_2, ..., x_n$ are sample values independently drawn from population with mean m and variance σ^2. Consider estimates

$$Y_n = \frac{X_1 + X_2 + ... + X_n}{n+1}, \quad Z_n = \frac{X_1 + 2X_2 + 3X_3 + ...nX_n}{n^2}$$

discuss whether they are unbiased, consistent for m. What is the efficiency of Y_n over Z_n?

3. Let X_1, X_2, X_3 and X_4 be independent random variables such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2, i = 1, 2, 3, 4$. If $Y = \frac{X_1 + X_2 + X_3 + X_4}{4}$, $Z = \frac{X_1 + X_2 + X_3 + X_4}{5}$ and $T = \frac{X_1 + 2X_2 - X_3 - X_4}{4}$, examin whether Y, Z and T are unbiased estimators of μ? What is the efficiency of Y relative to Z?

4. Let x_1, x_2, x_3, x_4 be a random sample from a $N(\mu, \sigma^2)$ population. Find the efficiency of

$$T = \frac{1}{7}(x_1 + 3x_2 + 2x_3 + x_4) \text{ relative to}$$

$$\overline{X} = \frac{1}{4}\sum_{i=1}^{n} x_i$$

which is relatively more efficient? Why?

5. What is an efficient estimator? If T_1 and T_2 are both efficient estimators with variance V and if $T = \frac{1}{2}(T_1 + T_2)$, show that variance of T is $\left(V/_2\right)(1+\rho)$ where ρ is the coefficient of correlation between T_1 and T_2 deduce that $\rho = 1$ and T is also efficient.

6. If T and T' be two consistent estimators of which T is the most efficient, prove that correlation coefficient between them is

$$\sqrt{\frac{V(T)}{V(T')}}$$, where $V(T)$ and $V(T')$ are the variances of T and T' respectively.

7. Examine the unbiasedness of the estimate $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$ where μ is known and σ^2 is the population variance in normal distribution.

7.14 Summary:

In this lesson the problem of point estimation and properties of a good estimator such as consistency, unbiasedness, bias and mean square error, efficiency and sufficiency are discussed. Derivations of sufficient statistics in Normal, Exponential, binomial and Poisson distributions are given. Some problems are solved to explain some of the properties and some problems are given to the students to solve on their own.

7.15 Technical Terms:

Consistency

Unbiasedness

Efficiency

Sufficiency of estimators

Neyman-Factorization Theorem.

Lesson Writer

V. Ramakrishna

Lesson 8

METHODS OF ESTIMATION

Objective:

After studying the lesson the students will be conversant with some methods of point estimation and interval estimation.

Structure of the Lesson:

8.1 Introduction:

The methods to estimate the parameters of the population using its representative called sample are many in number but we have in our syllabi, method of moments and method of maximum likelihood estimation. In this chapter we discussed the two methods of estimation, confidence interval and limits.

8.2 Methods of Moments:

This method is introduced by Karl Pearson and it is simplest method of finding estimator to the parameter by using the moments.

Let $f\left(x,\theta_1,\theta_2,..........,\theta_K\right)$ be the density function of the parent population with K parameters $\theta_1,\theta_2,..........,\theta_K$.

If μ_r^1 denotes the r^{th} moment about origin, then

$$\mu_r^1=\int x^r f\left(x;\theta_1,\theta_2,..........,\theta_K\right)dx,\ \left(r=1,2,............,K\right).$$

In general, $\mu_1^1,\mu_2^1,............,\mu_K^1$ will be functions of the parameters $\theta_1,\theta_2,..........,\theta_K$.

Let $x_i, i = 1, 2, ..., n$ be a random sample of size n from a given population. This method consists in solving the K equations for $\theta_1, \theta_2, ..., \theta_K$ interms of $\mu_1^1, \mu_2^1, ..., \mu_K^1$ and then replacing these moments $\mu_r^1, r = 1, 2, ...K$ by sample moments.

i.e. $$\hat{\theta}_i = \theta_i\left(\hat{\mu}_1^1, \hat{\mu}_2^1, ..., \hat{\mu}_K^1\right)$$

$$= \theta_i\left(m_1^1, m_2^1, ..., m_K^1\right), i = 1, 2, ...K$$

where m_i^1 is the i^{th} moment about origin in the sample.

Then by the method of moments $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K$ are the required estimators of $\theta_1, \theta_2, ..., \theta_K$ respectively.

Properties:

1. The sample moments are consistent estimators of the corresponding population moments.
2. It has been shown that under quite general conditions, the estimates obtained by the method of moments are asymptotically normal.
3. Generally the method of moments yields less efficient estimators than those obtained by maximum likelihood estimation. The estimators obtained by method of moments are identical with those given by method of maximum likelihood estimation if the probability density function is of exponential family.

Example: By the method of moments, find the estimators to the parameters in Normal population $N\left(\mu, \sigma^2\right)$.

Solution: From the given population μ is the arithmetic mean and σ^2 is the variance.

The population moments will be $\mu_1^1 = \mu$

$$\mu_2^1 = \sigma^2 + \mu^2.$$

If the sample moments with a sample size 'n' are m_1^1, m_2^1, then

$$m_1^1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$m_2^1 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 .$$

If we equate sample moments with population moments, we get

$$\bar{x} = \mu$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma^2 + \mu^2$$

Solving the above equations, we get $\hat{\mu} = \bar{x}$

$$\hat{\sigma}^2 = \frac{\sum {x_i}^2}{n} - \bar{x}^2 = s^2 .$$

Therefore, $\bar{x}$ and s^2 are the moment estimators for μ and σ^2 respectively.

8.3 Method of Maximum Likelihood Estimation:

This method was introduced by Prof. R.A. Fisher and later on developed by him in a series of papers.

The principle of maximum likelihood consists in finding an estimator for the unknown parameter θ which maximises the likelihood function $L(\theta)$ for variations in parameter.

We wish to find $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K\right)$ so that

$$L\left(\hat{\theta}\right) = \text{Sup } L(\theta) \; \forall \; \theta .$$

Thus if there exists a function $\hat{\theta} = \hat{\theta}(x_1, x_2, ..., x_n)$ of sample values which maximises L for variations in θ, then $\hat{\theta}$ is to be taken as an estimator of θ. $\hat{\theta}$ is called maximum likelihood estimator (M.L.E.) of θ.

Thus $\hat{\theta}$ is the solution, if any, of

$$\frac{\partial L}{\partial \theta} = 0 \text{ and } \frac{\partial^2 L}{\partial \theta^2} < 0 .$$

Since $L > 0$, and $\log L$ is a non-decreasing function of L, L and $\log L$ attain their extreme values (maxima or minima) at the same value of $\hat{\theta}$. The above two equations can be written as

$$\frac{\partial \log L}{\partial \theta} = 0 \text{ and } \frac{\partial^2 \log L}{\partial \theta^2} < 0 ,$$

a form which is much more convenient from practical point of view.

8.4 Properties of M.L.Estimators:

The following are the regularity conditions:

a) The first and second order derivatives $\frac{\partial \log L}{\partial \theta}, \frac{\partial^2 \log L}{\partial \theta^2}$ exist and are continuous functions of θ in a range R.

b) Third order derivative $\frac{\partial^3 \log L}{\partial \theta^3}$ exists such that $\left|\frac{\partial^3 \log L}{\partial \theta^3}\right| < M(x)$, where $E[M(x)] < K$, a positive quantity.

c) For every θ in R

$$E\left(\frac{-\partial^2 \log L}{\partial \theta^2}\right) = \int_{-\infty}^{\infty}\left(-\frac{\partial^2 \log L}{\partial \theta^2}\right) Ldx = I(\theta)$$

is finite and non-zero.

d) The range of integration is independent of θ. But if the range of integration depends on θ, then $f(x,\theta)$ vanishes at the extremes depending on θ.

Properties:

1. M.L. estimators are consistent.
2. M.L. estimators need not be unbiased.
3. ML estimators are asymptotically normally distributed.
4. If M.L.E. exists it is asymptotically most efficient.
5. If a sufficient estimator exists, it is a function of the maximum likelehood estimator.

8.5 Interval Estimation:

Let $x_i (i = 1, 2, ..., n)$ be a random sample of n observations from a population involving a single parameter θ. Let $f(x,\theta)$ be the probability function of the parent population from which the sample is drawn.

Let $t = t(x_1, x_2, ..., x_n)$ be a function of the sample values and an estimate of the population parameter θ, with the sampling distribution given by $g(t,\theta)$.

Neyman introduced the technique of confidence interval where in some reasonable probability statemets are made about the unknown parameter θ in the population. For a predetermined level of significance α (a small positive quantity), we have to determine two constants C_1 and C_2 such that $P(C_1 < \theta < C_2 | t) = 1 - \alpha$. The quantities C_1 and C_2 so determined, are known as confidence limits or fidicial limits. The unknown value of the population parameter is expected to lie in the interval (C_1, C_2) is called the confidence interval and $(1-\alpha)$ is called the confidence coefficient.

Let T_1 and T_2 be two statistics such that

$$P(T_1 > \theta) = \alpha_1 \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

and $$P(T_2 < \theta) = \alpha_2 \cdots\cdots\cdots\cdots\cdots (2)$$

where α_1 and α_2 are constants independent of θ.

The above equations combined together will give

$$P(T_1 < \theta < T_2) = 1 - \alpha ,$$

where $\alpha = \alpha_1 + \alpha_2$. Statistics T_1 and T_2 defined in (1) and (2) may be taken as C_1 and C_2.

8.6 Examples:

Example 1: The 95% confidence interval for the population Mean μ in normal population with known standard deviation σ in the case of large samples.

Solution: If we take a large sample from $N(\mu, \sigma^2)$, then

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

and $P(-1.96 \le Z \le 1.96) = 0.95$

(From Normal Probability Tables)

$$\Rightarrow P\left(-1.96 \le \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \le 1.96\right) = 0.95$$

$$\Rightarrow P\left(-1.96 \, \sigma/\sqrt{n} \le \bar{x} - \mu \le 1.96 \, \sigma/\sqrt{n}\right) = 0.95$$

Thus $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ are 95% confidence limits for the unknown parameter μ, the population mean and the interval $\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is called the 95% confidence interval.

Also $P(-2.58 \le Z \le 2.58) = 0.99$ (From Normal Probability Tables).

Hence, 99% confidence limits for μ are:

$$\bar{x} \pm 2.58\frac{\sigma}{\sqrt{n}}$$

and 99% confidence interval for μ is $\left(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}\right)$.

Example 2: Obtain $100(1-\alpha)\%$ confidence interval for the parameter σ^2 of the normal distribution.

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

Solution: Let $x_i \,(i = 1, 2, \ldots\ldots\ldots\ldots, n)$ be a random sample of size n from the density $f\left(x, \mu, \sigma^2\right)$ with known $\mu = \mu_0$ (say) and let $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \cdot$

Then $\frac{\sum(x_i - \mu_0)^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2(n) \cdot$

If we define χ^2_α, such that probability density function

$$P\left(\chi^2 > \chi^2_\alpha\right) = \int_{\chi^2_\alpha}^{\infty} P\left(\chi^2\right) d\chi^2 = \alpha,$$

where $P\left(\chi^2\right)$ is the p.d.f. of χ^2 distribution with n degrees of freedom. Then the required interval is given by

$$P\left(\chi^2_{\left(1-\alpha/2\right)} \le \chi^2 \le \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow P\left(\chi^2_{\left(1-\alpha/2\right)} \le \frac{ns^2}{\sigma^2} \le \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$\frac{ns^2}{\sigma^2} \le \chi^2_{\alpha/2} \Rightarrow \frac{ns^2}{\chi^2_{\alpha/2}} \le \sigma^2$$

$$\chi^2_{\left(1-\alpha/2\right)} \le \frac{ns^2}{\sigma^2} \Rightarrow \sigma^2 \le \frac{ns^2}{\chi^2_{\left(1-\alpha/2\right)}}$$

$$P\left(\frac{ns^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{ns^2}{\chi^2_{\left(1-\alpha/2\right)}}\right) = 1-\alpha$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are obtained from $P\left(\chi^2\right)$ with n degrees of freedom.

The 95% confidence interval for σ^2 is

$$P\left(\frac{ns^2}{\chi^2_{0.025}} \le \sigma^2 \le \frac{ns^2}{\chi^2_{0.975}}\right) = 0.95$$

Example 3: In random sampling from normal population $N\left(\mu, \sigma^2\right)$ find the maximum likelihood estimators for

(i) μ when σ^2 is known

(ii) σ^2 when μ is known

(iii) The simultaneous estimation of μ and σ^2.

Solution: $X \sim N\left(\mu, \sigma^2\right)$, then $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ is a random sample from normal population and the likelihood function is given by $L = \prod_{i=1}^{n} f\left(x_i, \mu, \sigma^2\right)$.

$$L = \prod_{i=1}^{n}\left[\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{\frac{-1}{2\sigma^2}\left(x_i-\mu\right)^2\right\}\right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^{n}\left(x_i-\mu\right)^2 \Big/ 2\sigma^2\right\}$$

$$\log L = \frac{-n}{2}\log\left(2\pi\right)-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2$$

Case (i) When σ^2 is known,

the likelihood equation for estimating μ is

$$\frac{\partial \log L}{\partial \mu} = 0$$

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

$$\sum_{i=1}^{n}\left(x_i-\mu\right)=0 \Rightarrow \hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x}$$

Case (ii) When μ is known, the likelihood equation for estimating σ^2 is

$$\frac{\partial}{\partial\sigma^2}\log L = 0 \Rightarrow -\frac{n}{2}\times\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(x_i-\mu\right)^2=0$$

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

$$\Rightarrow \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2 .$$

Case (iii) The likelihood equations for simultaneous estimation of μ and σ^2 are:

$$\frac{\partial \log L}{\partial \mu}=0 \text{ and } \frac{\partial \log L}{\partial \sigma^2}=0$$

$$\hat{\mu}=\overline{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum\left(x_i - \hat{\mu}\right)^2$$

$$= \frac{1}{n}\sum\left(x_i - \overline{x}\right)^2 = s^2$$ the sample variance.

Example 4: Find the maximum likelihood estimate for the parameter λ of a Poisson distribution on the basis of a sample of size n.

Solution: The probability function of Poisson distribution with parameter λ is given by

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

Likelihood function of random sample $x_1, x_2, ..., x_n$ of n observations from this population is

$$L = \prod_{i=1}^{n} P\left(X = x_i\right) = \frac{e^{-n\lambda}\lambda^{\sum\limits_{i=1}^{n} x_i}}{x_1!x_2!..........x_n!}$$

$$\log\ L = -n\lambda + n\overline{x}\log\lambda - \sum_{i=1}^{n}\log\left(x_i!\right)$$

The likelihood equation for estimating λ is

$$\frac{\partial \log L}{\partial \lambda} = 0 \Rightarrow -n + \frac{n\overline{x}}{\lambda} = 0$$

$$\therefore \hat{\lambda} = \overline{x}$$

Thus the M.L.E. for λ is the sample mean $\overline{x}$.

Example 5: Obtain the maximum likelihood estimated of P in the case of a sample observations $x_1, x_2, ..., x_n$ of size n drawn from binomial distribution $B\left(n, P\right)$.

Solution: If $x_1, x_2, ..., x_n$ is a random sample drawn from binomial distribution the likelihood function of sample $x_1, x_2, ..., x_n$ is

$$L = \prod_{i=1}^{n}\binom{n}{x_i}p^{\sum x_i}q^{\sum\left(n - x_i\right)}$$ where $q = 1 - p$.

$$\log L = \sum_{i=1}^{n} \log\binom{n}{x_i} + \sum x_i \log P + \left(n^2 - \sum x_i\right) \log q$$

$$= \sum_{i=1}^{n} \log\binom{n}{x_i} + \sum x_i \log P + \left(n^2 - \sum x_i\right) \log (1-P).$$

Consider $\frac{\partial \log L}{\partial P} = 0 \Rightarrow \frac{\sum_{i=1}^{n} x_i}{P} + \frac{\left(n^2 - \sum_{i=1}^{n} x_i\right)}{(1-P)}(-1) = 0$

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

$$\sum_{i=1}^{n} x_i - P\sum_{i=1}^{n} x_i - n^2 P + P\sum_{i=1}^{n} x_i = 0$$

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

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

$$P = \frac{\sum_{i=1}^{n} x_i}{n^2} = \frac{\overline{x}}{n}$$

$$\hat{P} = \frac{\overline{x}}{n}$$

Thus $\overline{x}/n$ is the M.L.E. for the parameter P in binomial distribution.

Example 6: Obtain maximum likelihood estimate of θ in $f(x,\theta) = (1+\theta)x^{\theta},\ 0 < x < 1$ based on a sample of size n.

Solution: The likelihood function of sample observations $x_1, x_2, \ldots\ldots\ldots, x_n$ is

$$L(x,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

$$= (1+\theta)^n \left(\prod_{i=1}^{n} x_i \right)^{\theta}$$

$$\log L = n \log (1+\theta) + \theta \sum_{i=1}^{n} \log x_i$$

$$\frac{\partial \log L}{\partial \theta} = \frac{n}{1+\theta} + \sum_{i=1}^{n} \log x_i = 0$$

$$\Rightarrow n + \sum_{i=1}^{n} \log x_i + \theta \sum_{i=1}^{n} \log x_i = 0$$

$$\hat{\theta} = \frac{-n}{\sum_{i=1}^{n} \log x_i} - 1 \cdot$$

Example 7: If $x_1, x_2, \ldots\ldots\ldots, x_n$ is a random sample of size n from a uniform population with p.d.f.

$$f(x,\theta) = 1,\ \theta - \frac{1}{2} \le x \le \theta + \frac{1}{2},\ -\infty < \theta < \infty \cdot$$

Obtain M.L.E. for θ.

Solution: The likelihood function of sample $x_1, x_2, \ldots\ldots\ldots, x_n$ is

$$L = L(\theta, x_1, x_2, \ldots\ldots\ldots, x_n) = 1,\ \theta - \frac{1}{2} \le x_i \le \theta + \frac{1}{2} \cdot$$

Since likelihood L is a constant, then the first order or second order derivatives w.r.t. θ are zeroes.

$\therefore$ Principle of maxima and minima fails to give MLE for θ.

As the range of the random variable involves the parameter θ, we find the MLE using order statistics.

If $x_{(1)}, x_{(2)}, \ldots\ldots\ldots, x_{(n)}$ is the ordered sample, then

$$\theta - \frac{1}{2} \le X_{(1)} \le X_{(2)} \le \cdots\cdots\cdots\cdots \le X_{(n)} \le \theta + \frac{1}{2} \cdot$$

Thus L attains maximum if

$$\theta-\frac{1}{2}\le x_{(1)} \text{ and } x_{(n)}\le\theta+\frac{1}{2}$$

$$\Rightarrow\theta\le x_{(1)}+\frac{1}{2} \text{ and } x_{(n)}-\frac{1}{2}\le\theta\cdot$$

Every statistic $t=t\left(x_1,x_2,\text{..........},x_n\right)$ such that

$$x_{(n)}-\frac{1}{2}\le t\left(x_1,x_2,\text{..........},x_n\right)\le x_{(1)}+\frac{1}{2} \text{ provides an M.L.E. for } \theta.$$

Example 8: Let $x_1,x_2,\text{..........},x_n$ be a random sample from the distribution with probability function.

$$f\left(x,\theta\right)=\frac{1}{\theta}\exp\left(-x/\theta\right),\ x>0\cdot$$

Find the maximum likelihood estimator of θ.

Solution: The likelihood equation of the sample observations $x_1,x_2,\text{..........},x_n$ is

$$L=f\left(x_1,x_2,\text{..........},x_n\mid\theta\right)$$

$$=f\left(x_1,\theta\right),f\left(x_2,\theta\right),\text{............},f\left(x_n,\theta\right)$$

$$=\frac{1}{\theta}\exp\left(-\frac{x_1}{\theta}\right)\frac{1}{\theta}\exp\left(-\frac{x_2}{\theta}\right)\text{..........}\frac{1}{\theta}\exp\left(-\frac{x_n}{\theta}\right)$$

$$=\left(\frac{1}{\theta}\right)^n\exp\left(-\frac{\sum x_i}{\theta}\right)=\theta^{-n}\exp\left(-\frac{\sum x_i}{\theta}\right)$$

$$\log L=-n\log\theta+\left(-\frac{\sum x_i}{\theta}\right)$$

$$\frac{\partial\log L}{\partial\theta}=0\Rightarrow\frac{-n}{\theta}-\left(-\frac{\sum x_i}{\theta^2}\right)=0$$

$$-\frac{n}{\theta}+\frac{\sum x_i}{\theta^2}=0$$

$$+\frac{n}{\cancel{\theta}}=+\frac{\sum x_i}{\theta^{\cancel{2}}}$$

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

$$\hat{\theta}=\overline{x}$$

$\therefore$ The maximum likelihood estimator of θ is the sample mean.

Example 9: Find moment estimators for the parameters of Normal Population.

Solution: Let $(x_1, x_2, \ldots\ldots\ldots, x_n)$ be a random sample of size n drawn from the Normal Population with

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

Let $m_1^1=\frac{1}{n}\sum_{i=1}^{n} x_i$ and $m_2=\frac{1}{n}\sum_{i=1}^{n} x_i^2$

be the first and second moments about origin, of the sample.

For the normal population we have

$$\mu_1^1=E(X)$$
$$=\mu$$

and $\mu_2=\mu_2^1-\mu_1^{1^2}$

$$=V(X)=\sigma^2$$

$$\Rightarrow \mu=\mu_1^1 \text{ and } \sigma^2=\mu_2^1-\mu_1^{1^2}$$

$$=\mu_2^1-\mu^2$$

Now replacing μ_1^1 and μ_2^1 by respective sample moments m_1^1 and m_2^1,

We obtain $\hat{\mu} = m_1^1 = \overline{x}$

$$\text{and } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\overline{x}\right)^2$$

$$= s^2$$

$\therefore$ $\overline{x}$ is the moment estimator for the parameter μ and s^2 is the moment estimator for σ^2.

8.7 Model Questions and Exercises:

1. Expalin the method of M.L.E. and obtain M.L.E. for mean parameter of Poisson distribution.
2. What are the different methods of estimation.
3. Describe the method of moments?
4. Obtain the moment estimators of the parameters μ and σ^2 of Normal Distribution, if a sample of size 'n' drawn randomly.
5. Obtain the M.L.E. of λ in the case of Poisson distribution and show that $\overline{x}$ is sufficient to estimate λ.
6. Using the statement of Factorization theorem obtain the sufficient statistics for μ, σ^2 in the case Normal Population.
7. Obtain the M.L.E. in Binomial Distribution $B(10,P)$.
8. Obtain 95% confidence interval for μ in the case of $N(\mu,10)$, if a sample of size 20 has been drawn from it.
9. Obtain the M.L.E. of θ in the case of exponential distribution $f(x,\theta) = \theta e^{-\theta x}, x > 0, \theta > 0$.
10. If $x_1, x_2, ..., x_n$ denote a random sample from a population with p.d.f. $f(x,\theta) = \theta x^{\theta-1}, \theta < x < 1$.

    Show that $Y = x_1, x_2, ..., x_n$ is a sufficient statistic for θ.
11. Examine if the following distribution admits a sufficient statistic for the parameter θ.

    $f(x,\theta) = (1+\theta)x^{\theta}, \ 0 \le x \le 1, \theta > 0$

12. Let $x_1, x_2, ..., x_n$ be a random sample from the distribution with p.d.f.

$$f(x,\theta) = \frac{1}{\theta}\exp\left(-x/\theta\right) \qquad \begin{array}{c} 0 < x < \infty \\ 0 < \theta < \infty \end{array}$$

Find the M.L.E. of θ

13. Obtain the M.L.E. for the parameter θ with the p.d.f.

$$f(x,\theta) = \theta^x (1-\theta)^{1-x},\ x = 0,1,\ 0 \le \theta \le 1 \cdot$$

14. Find the M.L.E. of θ for a random sample of size n from the distribution

$$f(x,\theta) = (\theta+1)x^{\theta} \qquad \begin{array}{c} 0 \le x \le 1 \\ \theta > -1 \end{array}$$

$= 0$, otherwise

15. Obtain the M.L.E. of θ based on a random sample of size n from the population with p.d.f.

(i) $f(x,\theta) = e^{-(x-\theta)},\ \theta \le x < \infty$

(ii) $f(x,\theta) = \theta x^{\theta-1},\ 0 < x < 1 \cdot$

Examine in each case, whether θ is unbiased.

8.8 Summary:

Two methods of estimation namely method of moments and M.L. estimation and the concept of Interval estimation are explained. A good number of examples are given to explain the above methods. Some model questions and exerises are given to the students to prepare and solve on their own.

8.9 Technical Terms:

Likelihood Function

Method of Moments

Maximum likelihood estimation

Interval Estimation.

Lesson Writer

V. Ramakrishna

Lesson 9

TESTING OF HYPOTHESIS

Objective:

After reading this lesson the students will be conversant with the concepts of null hypothesis, alternative hypothesis, critical region, two types of errors, level of significance and power of test. Neyman-Pearson lemma for testing a simple hypothesis against a simple alternative hypothesis.

Structure of the Lesson:

9.1 Introduction:

The purpose of hypothesis testing is to aid the researcher or administrator in reaching a decision concerning a population by examining a sample from that population. The theory of testing parametric statistical hypotheses was originally set forth by J. Neyman in 1928 and Karl Pearson in 1933. The basic concepts essential to an understanding of hypothesis testing problem and for obtaining a most powerful test are presented.

9.2 Basic Concepts:

A hypotesis may be defined simply as a statement about one or more populations. Researchers are concerned with two types of hypotheses research hypothesis and statistical hypothesis.

9.2.1 Research Hypothesis: The research hypothesis is the conjecture or supposition that motivates the research. It may be the result of years of observation on the part of the researcher.

Examples:

(1) A hospital administrator may hypothesize that the average length of stay of patients admitted to the hospital is five days.

(2) A public health nurse may hypothesize that a particular educational program will result in improved communication between nurse and patient.

(3) A doctor may hypothesize that a certain drug will be effective in 90 percent of the cases with which it is used.

By means of hypothesis testing one determines whether or not the above statements are compatible with available data. Research hypotheses lead directly to statistical hypotheses.

9.2.2 Statistical Hypothesis: Simple and Composite: A statistical hypothesis is an assertion or a statement abot the parameter (s) of one or more populations. In other words, a statistical hypothesis is an assertion about the distribution of one or more random variables. If the hypothesis completely specifies the distribution, it is called a simple hypothesis; otherwise it is called a composite hypothesis.

Examples: In normal distribution with parameters μ and σ, some examples of simple and composite hypotheses are given.

(1) A hypothesis of the form $\mu = \mu_0$, $\sigma = \sigma_0$ (where μ_0 and σ_0 are known values) is a simple hypothesis.

(2) If σ is known, then $\mu < 75$ $(\text{or } \mu > 75)$ is a composite hypothesis.

(3) If σ is unknown, then $\mu = 75$ is a composite hypothesis.

Hence, composite hypothesis is composed of a finite or infinite number of simple hypotheses. A simple hypothesis specifies a unique point in the parameter space Ω where as a composite hypothesis specifies more than one point in Ω.

9.2.3 Statistical Hypothesis: Null and Alternative: There are two statistical hypotheses involved in hypothesis testing. The first is the hypothesis to be tested, usually referred to as the null hypothesis and designated by the symbol H_0. The null hypothesis is sometimes referred to as a hypothesis of no difference, since it is a statement of agreement with conditions presumed to be true in the population of interest. In the testing process the null hypothesis is either rejected or accepted (not rejected). If the null hypothesis is not rejected, we will say that the data on which the test is based do not provide sufficient evidence to cause rejection. If the testing procedure leads to rejection, we will say that the data at hand are not compatible with the null hypothesis, but are supportive of some other hypothesis. This other hypothesis is known as the alternative hypothesis and designated by the symbol H_1. Both null and alternative hypotheses are framed before collecting the sample data from the population of interest and they are verified on the basis of a test of hypothesis.

Examples:

(1) Suppose that we want to answer the question: can we conclude that a certain population mean is not 75? The null hypothesis is $H_0 : \mu = 75$ and the alternative is $H_1 : \mu \neq 75$.

(2) Suppose we want to know if we could conclude that the population mean is greater than 75. Our hypotheses are $H_0 : \mu = 75,\ H_1 : \mu > 75$.

(3) If we want to know if we can conclude that the population mean is less than 75, the hypotheses are $H_0 : \mu = 75, H_1 : \mu < 75$.

(4) Two manufacturers making the same type of bulb as brand A and brand B, claim the superiority of their product over the other, interms of life of the bulb in hours. In such a cse, for testing the superiority of one brand over the other with respect to average life of bulb, we arrive at three possible statements.

(a) There is no difference in the average life of bulbs made by both the manufacturers. i.e., $\mu_A = \mu_B$ is a null hypothesis.

(b) The average life of brand A bulb is superior to that of brand B bulb i.e., $\mu_A > \mu_B$ is an alternative hypothesis.

(c) The average life of brand B bulb is superior to that of brand A bulb. i.e., $\mu_B > \mu_A$ is an alterntive hypothesis.

9.2.4. Test of a Hypothesis: A test of a statistical hypothesis is a rule which, when the experimental sample values have been obtained, leads to a decision to accept or to reject the hypothesis under consideration. So, test of a hypothesis arises only when we have to choose between two actions, say A_0 and A_1 corresponding to H_0 and H_1 respectively.

9.2.5. Two types of Errors: While testing a null hypothesis H_0 against an alternative hypothesis H_1, based on the sample data, we may arrive at two types of errors called type I error and type II error.

Type - I error: Rejecting H_0 when H_0 is true.

Type - II error: Accepting H_0 when H_1 is true (i.e., H_0 is false)

Examples:

(1) A student failed in the examination even though he wrote the examination very well, is a type I error. A student passed in the examination even though he wrote nothing, is a type II error.

(2) In quality control problems, we decide to accept or reject the lot after examining a sample from it. Here, type I error amounts to rejecting a lot when it is good, and type II error may be regarded as accepting a lot when it is bad.

9.2.6. Sample Space: The set of points representing the possible outcomes of an experiment is called the sample space of the experiment which is denoted by S.

9.2.7. Critical Region: If S is a sample space associated with the outcomes of an experiment, we divide the sample space into two disjoint regions $\overline{W}$ and W such that $\overline{W} \cup W = S$. The region W is that region where H_0 is rejected even though it is true, in accordance with a prescribed test. Then W is called the critical region or the rejection region of the test, and $\overline{W}$ is called the acceptance region of the test.

9.2.8. Probabilities of type I and type II errors: Let $(x_1, x_2, ..., x_n)$ be a random sample drawn from a population with parameter θ. To test a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative hypothesis $H_1 : \theta = \theta_1$, L be the likelihood function of a sample of observations with parameter θ, L_0 and L_1 are the likelihood functions under H_0 and H_1 respectively.

Now, symbolically we have

$$P\{\text{type I error}\} = \alpha$$

$$\Rightarrow P\{\text{rejecting } H_0 | H_0 \text{ is true}\} = \alpha$$

$$\Rightarrow P\{X \in W | L_0\} = \alpha$$ where W is the critical region.

Particularly, if X is a continuous random variable, we write the above as $\int_W L_0 dx = \alpha$.

Also

$$P\{\text{type II error}\} = \beta$$

$$\Rightarrow P\{\text{accepting } H_0 | H_1 \text{ is true}\} = \beta$$

$\Rightarrow P\left\{X \in \overline{W} \middle| L_1\right\} = \beta$, where $\overline{W}$ is the acceptance region.

$$\Rightarrow \int_{\overline{W}} L_1 \, dx = \beta .$$

Hence α and β are the probabilities of committing type I error and type II error respectively. Since α and β are the probabilities of committing wrong decisions, their values must be small. The problem of minimizing the two erros simultaneously is not possible. So, we fix α and minimize β in practice.

9.2.9. Power Function: The power function of a test of a statistical hypothesis H_0 against an alternative hypothesis H_1 is that function, defined for all distributions under consideration, which yields the probability that the sample point falls in the critical region W of the test, that is, a function that yields the probability of rejecting the hypothesis under consideration. When H_0 is true, power function yields the size α of the test. When H_1 is true, the power function yields the power $(1-\beta)$ of the test. The power $(1-\beta)$ is the probability of a correct decision. i.e., probability of rejecting H_0 when H_0 is false. Symbolically, $\int_{W} L_1 dx = 1-\beta \cdot$

9.2.10. Level of significance: The maximum value of the power function of the test when H_0 is true is called the level of significance of the test. In practice level of significance, size of the critical region, probability of committing type - I error are all treated equivalent. Generally α is chosen as 1% or 5%.

9.2.11. Best Critical Region (BCR): Let there be a class of test procedures for testing the same simple hypothesis against the same simple alternative hypothesis with equal probability of type I error. A test procedure whose powser is at least as great as the powers of all other test procedures in the class is called most powerfull test. The critical region associated with the most powerful test is called the best critical region. Neyman-Pearson Lemma provides a systematic method of determining a best critical region for testing a simple H_0 against a simple H_1.

9.3 Neyman - Pearson Lemma (NP Lemma):

Statement: Let $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ be a random sample from a distribution that has p.d.f. $f(x,\theta)$, L_0 and L_1 are likelihood functions of the sample observations $X = (x_1, \ldots\ldots\ldots, x_n)$ under H_0 and H_1 respectively. Let k be a positive number and W be a critical region of size α such that

$$W = \left\{ X \in S : \frac{L_0}{L_1} \le k \right\}$$

and $\overline{W} = \left\{ X \in S : \frac{L_0}{L_1} \ge k \right\}$,

where $\overline{W}$ is the acceptance region. i.e., $\overline{W} = S - W$.

Then W is a best critical region for testing a simple $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$.

Proof: We shall give the proof when the random variable is of continuous type. We are given

$$P\left(X \in W_1 \middle| H_0\right) = \int_W L_0 dX = \alpha \qquad \text{.........................} (1)$$

The power of the region W is

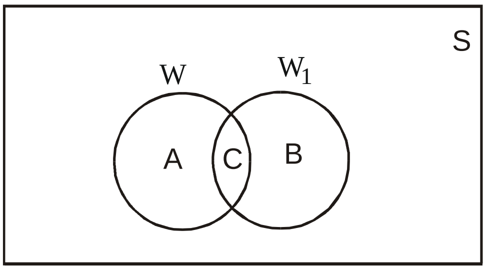
$$P\left(X \in W_1 \middle| H_1\right) = \int_W L_1 dX = 1 - \beta \qquad \text{.........................} (2)$$

In order to establish the lemma, we have to prove that there exists no other critical region of size α, which is more powerful then W. Let W_1 be another critical region of size α whose power is $1-\beta_1$ so that we have

$$P\left(X \in W_1 | H_0\right) = \int_{W_1} L_0 dX = \alpha \cdots\cdots\cdots\cdots\cdots (3)$$

and $$P\left(X \in W_1 | H_1\right) = \int_{W_1} L_1 dX = 1 - \beta_1 \cdots\cdots\cdots\cdots (4)$$

Now we have to prove that $(1-\beta) \ge (1-\beta_1)$.



Let $W = A \cup C$ and $W_1 = B \cup C$, we have

$$\int_W L_1 dX - \int_{W_1} L_1 dX = \int_C L_1 dX + \int_A L_1 dX - \int_C L_1 dX - \int_B L_1 dX$$

$$= \int_A L_1 dX - \int_B L_1 dX \qquad \text{.........................} (5)$$

However, by the hypothesis of the lemma, $L_1 \geq \frac{1}{k} L_0$ in the critical region W, and hence at each point of A; thus $\int_A L_1 dX \geq \frac{1}{k} \int_A L_0 dX$ .... (6)

But $L_1 \leq \frac{1}{k} L_0$ in the acceptance region $\overline{W}$, and hence at each point of B; accordingly,

$$\int_B L_1 dX \leq \frac{1}{k} \int_B L_0 dX \qquad \text{........................} (7)$$

Inequalities (6) and (7) imply that

$$\int_A L_1 dX - \int_B L_1 dX \geq \frac{1}{k} \int_A L_0 dX - \frac{1}{k} \int_B L_0 dX$$

and from equation (5) we obtain

$$\int_W L_1 dX - \int_{W_1} L_1 dX \geq \frac{1}{k} \left[\int_A L_0 dX - \int_B L_0 dX \right] \qquad \text{........................} (8)$$

However,

$$\int_A L_0 dX - \int_B L_0 dX = \int_A L_0 dX + \int_C L_0 dX - \int_C L_0 dX - \int_B L_0 dX$$

$$= \int_W L_0 dX - \int_{W_1} L_0 dX = \alpha - \alpha = 0 \qquad \text{........................} (9)$$

Substituting (9) in (8), we get

$$\int_{W} L_1 dX - \int_{W_1} L_1 dX \ \geq 0$$

i.e., $(1-\beta) \geq (1-\beta_1)$.

If the random variable is of discrete type, the proof is the same, with integration replaced by summation.

9.4 Problems:

Problem 1: If $x \geq 1$ is the critical region for testing $H_0 : \theta = 2$ agaist $H_1 : \theta = 1$ on the basis of a single observation from the population,

$$f(x, \theta) = \theta\, e^{-\theta x},\ 0 \leq x < \infty$$

Obtain the values of probabilities of type I and type II errors.

Solution: Given $W = \{\ x : x \geq 1\},\ H_0 : \theta = 2,\ H_1 : \theta = 1$

$$\alpha = P(\text{type I error}) = P(\text{rejecting } H_0 | H_0 \text{ is true})$$

$$= P\left[x \in W | H_0\right]$$

$$= P\left[x \geq 1 \mid \theta = 2\right]$$

$$= \int_1^{\infty} \left[f(x, \theta)\right]_{\theta = 2} dx$$

$$= 2\int_1^{\infty} e^{-2x} dx$$

$$= 2\left|\frac{e^{-2x}}{-2}\right|_1^{\infty}$$

$$= e^{-2}$$

$$= \frac{1}{e^2}$$

$$\beta = P(\text{type II error}) = P\left(\text{accepting } H_0 \middle| H_1 \text{ is true}\right)$$

$$= P\left[x \in \overline{W} \middle| H_1\right]$$

$$= P\left[x < 1 \middle| \theta = 1\right]$$

$$= \int_0^1 e^{-x}\,dx$$

$$= \left|\frac{e^{-x}}{-1}\right|_0^1$$

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

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

Problem 2: If $0.5 \le x$ is the critical region for testing $H_0 : \theta = 1$ against $H_1 : \theta = 2$, on the basis of a single observation drawn from a rectangular population where $x \in (0, \theta)$, obtain the values of α and β.

Solution: The p.d.f. of the rectangular distribution is $f(x,\theta) = \frac{1}{\theta}$ for $x \in (0,\theta)$.

Given $W = \{x : 0.5 \le x\},\ \overline{W} = \{x : x < 0.5\},\ H_0 : \theta = 1,\ H_1 : \theta = 2$

$$\alpha = P(\text{type I error}) = P\left(\text{rejecting } H_0 \middle| H_0 \text{ is true}\right)$$

$$= P\left\{x \in W \mid H_0\right\}$$

$= P\left\{x \ge 0.5 \mid \theta = 1\right\}$ for $0 < x < \theta$ and $\theta = 1$

$$= \int_{0.5}^{1} \left[f(x,\theta)\right]_{\theta=1}^{dx}$$

$$= \int_{0.5}^{1} 1 \; dx$$

$$= |x|_{0.5}^{1}$$

$$= 0.5$$

$$\beta = P(\text{type II error})$$

$$= P(\text{accepting } H_0 \mid H_1 \text{ is true})$$

$$= P\left(x \in \overline{W} \mid H_1\right)$$

$$= P\{x < 0.5 \mid \theta = 2\}$$

$$= \int_{0}^{0.5} \left[f(x,\theta)\right]_{\theta=2} dx$$

$$= \int_{0}^{0.5} \frac{1}{2} \, dx$$

$$= \frac{1}{2} |x|_{0}^{0.5}$$

$$= \frac{1}{2}(0.5)$$

$$= 0.25$$

Problem 3: Let p be the probability that a coin will fall head in a single toss. In order to test $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{3}{4}$, the coin is tossed 5 times and H_0 is rejected if the head appears more than 3 times. Find the size and power of the test.

Solution: Let x be the number of heads appeared when the coin is tossed 5 times and p be the probability of getting a head at each toss, then x follows binomial distribution with parameters n = 5 and p.

$\therefore$ The probability mass function is

$$P(x) = {}^{n}C_{x}\, p^{x}\, (1-p)^{n-x} \text{ for } x = 0,1,2,3,4,5$$

$$= {}^{5}C_{x}\, p^{x}\, (1-p)^{5-x}$$

$$= {}^{5}C_{x} \left(\frac{1}{2}\right)^{x} \left(1-\frac{1}{2}\right)^{5-x} \text{ When } H_0 \text{ is true.}$$

$$= {}^{5}C_{x} \left(\frac{1}{2}\right)^{5} \text{ When } H_0 \text{ is true.}$$

Similarly, $P(x) = {}^{5}C_{x} \left(\frac{3}{4}\right)^{x} \left(1-\frac{3}{4}\right)^{5-x}$ When H_1 is true

$$= {}^{5}C_{x} \left(\frac{3}{4}\right)^{x} \left(\frac{1}{4}\right)^{5-x}$$

$$= {}^{5}C_{x}\, \frac{3^{x}}{4^{x}}\, \frac{1^{5-x}}{4^{5-x}}$$

$$= {}^{5}C_{x}\, \frac{3^{x}}{4^{5}}$$

$$= \frac{1}{4^{5}}\, {}^{5}C_{x}\ 3^{x} \text{ when } H_1 \text{ is true.}$$

H_0 is rejected if the head appears more then 3 times in 5 tosses. i.e., the critical region is $W = \{x = 4,5\}$.

Size of the test $= \alpha = P(\text{type I error})$

$$\alpha = P\left(\text{rejecting } H_0 \middle| H_0 \text{ is true}\right)$$

$$= P\left\{x = 4,5 \mid 5_{C_x} \left(\frac{1}{2}\right)^{5}\right\}$$

$$= {}^{5}C_{4}\left(\frac{1}{2}\right)^{5} + {}^{5}C_{5}\left(\frac{1}{2}\right)^{5}$$

$$= 5\left(\frac{1}{2}\right)^5 + 1\left(\frac{1}{2}\right)^5$$

$$= \frac{1}{2^5}(5+1)$$

$$= \frac{6}{32}$$

$$= 0.1875$$

Power of the test $= 1 - \beta$

$$1-\beta = P\left(\text{rejecting } H_0 \mid H_1 \text{ is true}\right)$$

$$= P\left\{x = 4, 5 \mid \frac{1}{4^5}\, {}^5C_x 3^x\right\}$$

$$= \frac{1}{4^5}\, {}^5C_4 3^4 + \frac{1}{4^5}\, {}^5C_5 3^5$$

$$= \frac{1}{4^5}\left(5\times 81 + 1\times 243\right) = \frac{405+243}{1024}$$

$= 0.6328$

Problem 4: Let $X \sim N(\mu, 4)$. To test $H_0 : \mu = -1$ against $H_1 : \mu = 1$ based on a sampel of size $n = 10$ drawn from this population, the critical region used is $x_1 + 2x_2 + 3x_3 + ... + 10x_{10} \geq 0$. Find the size and power of the test.

Solution: Given that $X \sim N(\mu, 4)$, $n = 10$

To test $H_0 : \mu = -1$ against $H_1 : \mu = 1$ the critical region is

$$W = \{x_1 + 2x_2 + 3x_3 + ... + 10x_{10} \geq 0\}.$$

Let $U = x_1 + 2x_2 + 3x_3 + ... + 10x_{10}$ where each of the observations x_i are independent normal variates, then U is also a normal variate with mean

$$E(U) = E(x_1) + 2E(x_2) + 3E(x_3) + ... + 10E(x_{10})$$

$= \mu + 2\mu + 3\mu + ... + 10\mu$ since $E(x_i) = \mu$

$= \mu(1 + 2 + 3 + ... + 10)$

$= 55\,\mu$ since $n(n+1)/2 = 55$

$$V(U) = V(x_1) + 2^2 V(x_2) + 3^2 V(x_3) + ... + 10^2 V(x_{10})$$

$= \sigma^2 + 2^2\sigma^2 + 3^2\sigma^2 + ... + 10^2\sigma^2$ since $V(x_i) = \sigma^2$

$= 4\left(1^2 + 2^2 + 3^2 + ... + 10^2\right)$ since $\sigma^2 = 4$

$= 4 \times 385$ since $n(n+1)(2n+1)/6 = 385$

$= 1540$

Now, $E(U) = 55\mu$ and $V(U) = 1540$

The standard normal variate is

$$Z = \frac{U - E(U)}{\sqrt{V(U)}}$$

$$= \frac{U - 55\mu}{\sqrt{1540}}$$

$$= \frac{U - 55\mu}{39.24}$$

Now, the critical region is expressed as $W = \{U \geq 0\}$.

The size of the test $= \alpha = P(\text{type I error})$

$$\alpha = P(\text{rejecting } H_0 | H_0 \text{ is true})$$

$$= P(U \geq 0 | H_0 \text{ is true})$$

$$= P(39.24z + 55\mu \geq 0 \mid \mu = -1)$$

$$= P(39.24z - 55 \geq 0)$$

$$= P\left(Z \geq \frac{55}{39.24}\right)$$

$$= P\left(Z \geq 1.4\right)$$

$= 0.0808$ observed from standard normal distribution table.

The power of the test $= 1 - \beta$

$$1 - \beta = P\left(\text{rejecting } H_0 \,\middle|\, H_1 \text{ is true}\right)$$

$$= P\left(39.24z + 55\mu \geq 0 \,\middle|\, \mu = 1\right)$$

$$= P\left(39.24z + 55 \geq 0\right)$$

$$= P\left(Z \geq \frac{-55}{39.24}\right)$$

$$= P\left(z \geq -1.4\right)$$

= 0.9192 observed from standard normal distribution table.

Problem 5: Show that the power of a best critical region for testing a simple hypothesis against a simple alternative hypothesis is never less than its size.

Solution: Let W be a BCR of size α for testing a simple hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$. Let L_0 and L_1 are the likelihood functions under H_0 and H_1 respectively. Then

$$\int_W L_0 dX = \alpha \qquad \text{......................... (1)}$$

By Neyman - Pearson Lemma we have,

$\frac{L_0}{L_1} \leq k$ within the BCR W.

i.e., $k\, L_1 \geq L_0$

$$k \int_W L_1 dX \geq \int_W L_0 dX$$

$$k(1 - \beta) \geq \alpha \qquad \text{......................... (2)}$$

Again we have

$\frac{L_0}{L_1} \geq k$ within acceptance region $\overline{W}$.

i.e., $k\, L_1 \leq L_0$

$k \int_{\overline{W}} L_1\, dX \leq \int_{\overline{W}} L_0 dX$

$k\beta \leq (1-\alpha)$

$(1-\alpha) \geq k\beta$ .... (3)

From (2) and (3) we get

$k(1-\beta)(1-\alpha) \geq k\, \alpha\beta$

$1-\beta-\alpha+\alpha\beta-\alpha\beta \geq 0$

$(1-\beta) \geq \alpha$.

9.5 Exercises:

1. For testing the null hypothesis $H_0 : \theta = 1$ against $H_1 : \theta = 2$, a sample of size one observation is drawn from a rectangular population $f(x,\theta) = \frac{1}{\theta}$ where $x \in (0,\theta)$. What would be the sizes of type I and type II errors if you choose $1 \leq x \leq 1.5$ as the critical region.

2. On the basis of a single observation from the population whose p.d.f. is $f(x,\theta) = \theta e^{-\theta x}$, $0 \leq x < \infty$; we have to test $H_0 : \theta = 1$ against $H_1 : \theta = 4$ and reject H_0 when $x > 3$. Find the size of the critical region and power of the test.

3. Let X has the p.d.f. $f(x,p) = p^x (1-p)^{1-x}$, $x = 0,1$. To test $H_0 : p = \frac{1}{4}$ against $H_1 : p < \frac{1}{4}$ based on a random sample of 10 observations $x_1, x_2, ..., x_{10}$, the critical region is given by $\sum_{i=1}^{10} x_i \leq 1$. Find the power function of this test.

4. Let X has the p.d.f. $f(x,\lambda) = e^{-\lambda} \lambda^x / x!$, $x = 0,1,..............$ To test $H_0 : \lambda = 1$ against $H_1 : \lambda < 1$ based on a random sample of size 10, the critical region is $X_1 + X_2 + ... + X_{10} \le 4$. Find an expression for the power function of this test.

5. To test $H_0 : \mu = 1$ against $H_1 : \mu = 2$ based on a random sample of 10 observations drawn from $N(\mu, 4)$, the critical region used is $x_1 + 2x_2 + 3x_3 + ... + 10x_{10} \ge 0$. Find the size and power of the test.

9.6 Summary:

The basic concepts that we come across in hypothesis testing problem are explained. The statement and proof of the Neyman - Pearson Lemma for testing a simple hypothesis against a simple alternative hypothesis are given. Some problems related to probabilities of type I and type II errors, and power of the test are sovled while some problems are left over to the students as exercises to try on their own.

9.7 Technical Terms:

Statistical hypothesis

Type I error

Type II error

Critical region

Level of significance

Power function.

Lesson Writer

C.V. RAO

Lesson 10

ILLUSTRATIONS ON N - P LEMMA

Objective:

After studying this lesson the students will have clear comprehension on the practical utility of Neyman - Pearson Lemma and will be able to derive the best critical regions of tests in binomial, Poisson, exponential and normal (for mean with known S.D.) distributions.

Structure of the Lesson:

10.1 Introduction:

The statement and proof of the Neyman - Pearson (N - P) lemma was given in Lesson 9. The N - P lemma provides the best critical region (BCR) for testing a simple null hypothesis H_0 against a simple alternative hypothesis H_1. A test based on BCR is called a most powerful (MP) test. Using N - P lemma, best critical regions are obtained for testing of hypotheses problems in binomial distribution, Poisson distribution, exponential distribution and normal distribution (with known standard deviation).

In deriving the BCR, one can use any one of the two inequalities either $\frac{L_1}{L_0} \geq K$ or $\frac{L_0}{L_1} \leq K$, where L_0 and L_1 are the likelihood functions under H_0 and H_1 respectively, and K is a positive constant.

10.2 Example in Binomial Distribution:

Example 1: Obtain the best critical region for testing $H_0 : p = p_0$ against $H_1 : p = p_1$ for the parameter p in binomial distribution based on a random sample of size n.

Solution: Let $x_1, x_2, \ldots\ldots\ldots, x_n$ be a random sample drawn from a binomial population with probability mass function

$$P(X=x) = {}^{n}C_{x}\, p^{x}\,(1-p)^{n-x} \;\; ; x = 0,1,2,........., n$$

The likelihood funtion L is

$$L = P(x_1)\, P(x_2) ....P(x_n)$$

$$\Rightarrow L = \left(\pi_{i=1}^{n}\, {}^{n}C_{x_i}\right) p^{\sum_{i=1}^{n} x_i} \,(1-p)^{\sum_{i=1}^{n}(n-x_i)}$$

$$\Rightarrow L_0 = \left(\pi_{i=1}^{n}\, {}^{n}C_{x_i}\right) p_0^{\sum_{i=1}^{n} x_i} \,(1-p_0)^{\sum_{i=1}^{n}(n-x_i)} \quad \text{under } H_0$$

and $$L_1 = \left(\pi_{i=1}^{n}\, {}^{n}C_{x_i}\right) p_1^{\sum_{i=1}^{n} x_i} \,(1-p_1)^{\sum_{i=1}^{n}(n-x_i)} \quad \text{under } H_1.$$

The best critical region using N - P lemma is obtained by considering $\frac{L_1}{L_0} \geq K,$ where K is any positive constant.

Now, $\frac{L_1}{L_0} \geq K$

$$\Rightarrow \frac{\left(\pi_{i=1}^{n}\, {}^{n}C_{x_i}\right) p_1^{\sum_{i=1}^{n} x_i} \,(1-p_1)^{\sum_{i=1}^{n}(n-x_i)}}{\left(\pi_{i=1}^{n}\, {}^{n}C_{x_i}\right) p_0^{\sum_{i=1}^{n} x_i} \,(1-p_0)^{\sum_{i=1}^{n}(n-x_i)}} \geq K$$

$$\Rightarrow \left(\frac{p_1}{p_0}\right)^{\sum_{i=1}^{n} x_i} \left(\frac{1-p_1}{1-p_0}\right)^{\sum_{i=1}^{n}(n-x_i)} \geq K$$

Taking logarithms on both sides

$$\Rightarrow \sum x_i \ \log\left(\frac{p_1}{p_0}\right) + \sum(n - x_i)\left(\frac{1-p_1}{1-p_0}\right) \ \geq \log \ K$$

$$\Rightarrow \sum x_i \ \log\left(\frac{p_1}{p_0}\right) + \left(n^2 - \sum x_i\right) \log\left(\frac{1-p_1}{1-p_0}\right) \geq \log K$$

$$\Rightarrow \sum x_i \left\{\log\left(\frac{p_1}{p_0}\right) - \log\left(\frac{1-p_1}{1-p_0}\right)\right\} + n^2 \log\left(\frac{1-p_1}{1-p_0}\right) \geq \log K$$

$$\Rightarrow \sum x_i \ \log\left[\frac{p_1/p_0}{(1-p_1)/(1-p_0)}\right] \geq \log \ K - n^2 \log\left(\frac{1-p_1}{1-p_0}\right)$$

$$\Rightarrow \sum x_i \ \log\left[\frac{p_1(1-p_0)}{p_0(1-p_1)}\right] \geq \log K - n^2 \log\left(\frac{1-p_1}{1-p_0}\right)$$

If $p_1 > p_0$, the best critical region is given by

$$W = \left\{(x_1, x_2, \ldots\ldots\ldots\ldots, x_n) : \sum_{i=1}^{n} x_i \ \geq \frac{\log K - n^2 \log\left(\frac{1-p_1}{1-p_0}\right)}{\log\left[\frac{p_1(1-p_0)}{p_0(1-p_1)}\right]}\right\}.$$

If $p_1 < p_0$, the BCR is given by

$$W = \left\{(x_1, x_2, \ldots\ldots\ldots\ldots, x_n) : \sum_{i=1}^{n} x_i \ \leq \frac{\log \ K - n^2 \log\left(\frac{1-p_1}{1-p_0}\right)}{\log\left[\frac{p_1(1-p_0)}{p_0(1-p_1)}\right]}\right\}$$

10.3 Examples in Poisson Distribution:

Example 2 : Derive the best critical region for testing $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda = \lambda_1$ for the parameter λ in Poisson distribution based on a random sample of size n.

Solution: Let $x_1, x_2, \ldots\ldots, x_n$ be a random sample drawn from a Poisson distribution with probability mass function

$$P(X=x)=\frac{e^{-\lambda}\,\lambda^{x}}{x!},\quad \begin{array}{l} x=0,1,2,\ldots\ldots\ldots\ldots \\ \lambda>0. \end{array}$$

The likelihood function L is

$$L=P(x_1)\;P(x_2)\ldots\ldots\ldots\ldots P(x_n)$$

$$L=\frac{e^{-n\lambda}\;\lambda^{\sum_{i=1}^{n} x_i}}{\pi_{i=1}^{n} x_i\,!}$$

$$L_0=\frac{e^{-n\lambda_0}\;\lambda_0^{\sum x_i}}{\pi_{i=1}^{n} x_i\,!}\quad \text{under } H_0$$

$$L_1=\frac{e^{-n\lambda_1}\;\lambda_1^{\sum x_i}}{\pi_{i=1}^{n} x_i\,!}\quad \text{under } H_1$$

The BCR using N - P lemma is obtained by considering $\frac{L_1}{L_0}\ge K$ where K is any positive constant.

Now, $\frac{L_1}{L_0}\ge K$

$$\Rightarrow \frac{e^{-n\lambda_1}\,\lambda_1^{\sum x_i}\Big/\pi_{i=1}^{n} x_i!}{e^{-n\,\lambda_0}\,\lambda_0^{\sum x_i}\Big/\pi_{i=1}^{n} x_i!}\ \ge K$$

$$\Rightarrow e^{n\lambda_0-n\lambda_1}\;\frac{\lambda_1^{\sum x_i}}{\lambda_0^{\sum x_i}}\ \ge\ K$$

$$\Rightarrow e^{n(\lambda_0-\lambda_1)}\left(\frac{\lambda_1}{\lambda_0}\right)^{\sum x_i} \geq K$$

Taking logarithms on both sides

$$\Rightarrow n\left(\lambda_0-\lambda_1\right)+\sum x_i \log\left(\frac{\lambda_1}{\lambda_0}\right) \geq \log K$$

$$\Rightarrow \sum x_i \log\left(\frac{\lambda_1}{\lambda_0}\right)-n\left(\lambda_1-\lambda_0\right) \geq \log K$$

If $\lambda_1 > \lambda_0$

$$\sum x_i \log\left(\frac{\lambda_1}{\lambda_0}\right) \geq \log K+n\left(\lambda_1-\lambda_0\right)$$

$$\Rightarrow \sum x_i \geq \frac{\log K+n\left(\lambda_1-\lambda_0\right)}{\log\left(\frac{\lambda_1}{\lambda_0}\right)}.$$

$\therefore$ The BCR when $\lambda_1 > \lambda_0$ is given by

$$W=\left\{\left(x_1, x_2, \ldots\ldots\ldots, x_n\right): \sum_{i=1}^{n} x_i \geq \frac{\log K+n\left(\lambda_1-\lambda_0\right)}{\log\left(\frac{\lambda_1}{\lambda_0}\right)}\right\}.$$

If $\lambda_1 < \lambda_0$

$$\sum x_i \leq \frac{\log K+n\left(\lambda_1-\lambda_0\right)}{\log\left(\frac{\lambda_1}{\lambda_0}\right)}$$

$\therefore$ The BCR when $\lambda_1 < \lambda_0$ is given by

$$W=\left\{\left(x_1, \ldots\ldots, x_n\right): \sum_{i=1}^{n} x_i \leq \frac{\log K+n\left(\lambda_1-\lambda_0\right)}{\log\left(\frac{\lambda_1}{\lambda_0}\right)}\right\}.$$

Example 3: Given a sample of n observations $x_1, \ldots\ldots\ldots\ldots, x_n$, obtain the BCR for testing the simple hypothes is $H_0 : f(x) = e^{-1}/x!$, $x = 0,1,2,\ldots\ldots\ldots\ldots$ against the simple alternative $H_1 : f(x) = \left(\frac{1}{2}\right)^{x+1}$, $x = 0,1,2,\ldots\ldots\ldots$

Also obtain the size and power of the test.

Solution: The likelihood function under H_0 is

$$L_0 = \frac{e^{-n}}{\pi_{i=1}^{n}(x_i!)}$$

$$L_1 = \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^{\Sigma x_i}$$

Using N - P lemma the BCR is given by considering $\frac{L_0}{L_1} \le K$, where K is a poistive constant.

Now, $\frac{L_0}{L_1} \le K$

$$\Rightarrow \frac{e^{-n} \Big/ \pi_{i=1}^{n}(x_i!)}{\left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^{\Sigma x_i}} \le K$$

$$\Rightarrow \frac{e^{-n}\, 2^n\, 2^{\Sigma x_i}}{\pi_{i=1}^{n}(x_i!)} \le K$$

$$\Rightarrow \frac{\left(2e^{-1}\right)^n\, 2^{\Sigma x_i}}{\pi_{i=1}^{n}(x_i!)} \le K \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Taking 'log' on both sides

$$\left(\sum_{i=1}^{n} x_i\right) \log 2 - \log\left[\underset{i=1}{\overset{n}{\pi}} (x_i !)\right] \le \log K - n \log\left(2e^{-1}\right)$$

$\therefore$ The BCR is

$$W = \left\{(x_1,.........,x_n) : \left(\sum_{i=1}^{n} x_i\right) \log\ 2 - \log\left[\underset{i=1}{\overset{n}{\pi}} (x_i\ !)\right] \le C\right\}$$

where $C = \log K - n \log\left(2e^{-1}\right)$.

Consider the case $K = 1$ and $n = 1$. The above inequality (1) may be written as

$$\frac{\left(\frac{2}{e}\right) 2^{x_1}}{x_1\ !} \le 1$$

$$\Rightarrow \quad \frac{2^{x_1}}{x_1!} \le \frac{e}{2}$$

This inequality is satisfied by all the points in the set $W = \{x_1 : x_1 = 0, 3, 4, 5, ...\}$. Thus the size of the test when H_0 is true is

$$\alpha = P(x_1 \in C \mid H_0) = 1 - P(x_1 = 1, 2 \mid H_0)$$

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

$$= 1 - \frac{1.5}{2.7183} = 1 - 0.552 = 0.448$$

The power of the test when H_1 is true is

$$1 - \beta = P(x_1 \in C \mid H_1) = 1 - P(x_1 = 1, 2 \mid H_1)$$

$$= 1 - \left(\frac{1}{4} + \frac{1}{8}\right) = 0.625$$

10.4 Examples in Exponential Distribution:

Example 4: Derive the best critical region for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ for a sample of size n drawn from an exponential population with parameter θ.

Solution: Let $x_1, x_2, ..., x_n$ be a random sample drawn from an exponential population with p.d.f.

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

The likelihood funtion L is given by

$$L = f(x_1, \theta)\ f(x_2, \theta) \cdots\cdots\cdots\cdots f(x_n, \theta)$$

$$L = \theta^n\ e^{-\theta \sum_{i=1}^{n} x_i}$$

$$\Rightarrow L_0 = \theta_0^n\ e^{-\theta_0 \sum x_i} \ \text{ under } H_0$$

and $L_1 = \theta_1^n\ e^{-\theta_1 \sum x_i}$ under H_1

Using N - P lemma the BCR is obtained by considering $\frac{L_1}{L_0} > K$,

where K is a positive constant.

Now, $\frac{L_1}{L_0} \geq K$

$$\Rightarrow \frac{\theta_1^n\ e^{-\theta_1 \sum x_i}}{\theta_0^n\ e^{-\theta_0 \sum x_i}} \geq K$$

$$\Rightarrow \left(\frac{\theta_1}{\theta_0}\right)^n e^{(\theta_0 - \theta_1)\sum x_i} \geq K$$

Taking 'log' on both sides

$$\Rightarrow n \log\left(\frac{\theta_1}{\theta_0}\right) + \sum x_i\ (\theta_0 - \theta_1) \geq \log K$$

$$\Rightarrow \sum x_i \left(\theta_0 - \theta_1\right) \geq \log K - n \log\left(\frac{\theta_1}{\theta_0}\right)$$

$$\Rightarrow \sum x_i \left(\theta_1 - \theta_0\right) \leq n \log\left(\theta_1/\theta_0\right) - \log K$$

If $\theta_1 > \theta_0$

$$\sum x_i \leq \frac{n \log\left(\theta_1/\theta_0\right) - \log K}{\theta_1 - \theta_0}.$$

$\therefore$ The BCR when $\theta_1 > \theta_0$ is given by

$$W = \left\{\left(x_1,..........,x_n\right) : \sum_{i=1}^{n} x_i \leq \frac{n \log\left(\theta_1/\theta_0\right) - \log K}{\left(\theta_1 - \theta_0\right)}\right\}$$

If $\theta_1 < \theta_0$ the BCR is given by

$$W = \left\{\left(x_1,...........,x_n\right) : \sum_{i=1}^{n} x_i \geq \frac{n \log\left(\theta_1/\theta_0\right) - \log K}{\left(\theta_1 - \theta_0\right)}\right\}$$

Example 5: Obtain the BCR for testing the hypothesis $H_0 : \beta = \beta_0, \gamma = \gamma_0$ against the alternative hypothesis $H_1 : \beta = \beta_1, \gamma = \gamma_1$ based on a random sample of size n drawn from a population whose p.d.f. is given by

$$f\left(x,\beta,\gamma\right) = \beta e^{-\beta(x-\gamma)} \; ; x \geq \gamma$$

Solution: Given $f\left(x,\beta,\gamma\right) = \beta e^{-\beta(x-\gamma)} \; ; x \geq \gamma$

$$L = \underset{i=1}{\overset{n}{\pi}} f\left(x_i,\beta,\gamma\right)$$

$$L = \beta^n \, e^{-\beta \sum_{i=1}^{n}(x_i - \gamma)}$$

$$L_0 = \beta_0^n \, e^{-\beta_0 \sum(x_i - \gamma_0)} \quad \text{under } H_0$$

$$L_1 = \beta_1^n \, e^{-\beta_1 \sum(x_i - \gamma_1)} \quad \text{under } H_1.$$

Using N - P lemma the BCR is obtained by considering $\frac{L_1}{L_0} \geq K$, where K is a positive constant.

Now $\frac{L_1}{L_0} \geq K$

$$\Rightarrow \frac{\beta_1^n \; e^{-\beta_1 \sum(x_i - \gamma_1)}}{\beta_0^n \; e^{-\beta_0 \sum(x_i - \gamma_0)}} \geq K$$

$$\Rightarrow \left(\frac{\beta_1}{\beta_0}\right)^n e^{-\beta_1 \sum(x_i - \gamma_1) + \beta_0 \sum(x_i - \gamma_0)} \geq K$$

$$\Rightarrow \left(\frac{\beta_1}{\beta_0}\right)^n e^{-\beta_1 n\bar{x} + \beta_0 n\bar{x} + n\beta_1\gamma_1 - n\beta_0\gamma_0} \geq K$$

Taking 'log' on both sides

$$n \log\left(\frac{\beta_1}{\beta_0}\right) - n\,\bar{x}\,(\beta_1 - \beta_0) + n\beta_1\gamma_1 - n\beta_0\gamma_0 \geq \log K$$

$$\Rightarrow n\bar{x}\,(\beta_1 - \beta_0) \leq n \log\left(\frac{\beta_1}{\beta_0}\right) + n\beta_1\gamma_1 - n\beta_0\gamma_0 - \log K$$

$$\Rightarrow \bar{x} \leq \frac{\left\{\log\left(\frac{\beta_1}{\beta_0}\right) + \beta_1\gamma_1 - \beta_0\gamma_0 - \frac{\log K}{n}\right\}}{\beta_1 - \beta_0}$$ provided $\beta_1 > \beta_0$.

$\therefore$ The BCR when $\beta_1 > \beta_0$ is given by

$$W = \left\{ (x_1, \ldots\ldots\ldots, x_n) : \bar{x} \leq \frac{\left[\log\left(\frac{\beta_1}{\beta_0}\right) + \beta_1\gamma_1 - \beta_0\gamma_0 - \frac{\log K}{n}\right]}{(\beta_1 - \beta_0)} \right\}$$

If $\beta_1 < \beta_0$ the BCR is given by

$$W = \left\{ (x_1, ..., x_n) : \bar{x} \geq \frac{\left[\log\left(\frac{\beta_1}{\beta_0}\right) + \beta_1\gamma_1 - \beta_0\gamma_0 - \frac{\log K}{n} \right]}{(\beta_1 - \beta_0)} \right\}$$

10.5 Examples in normal Distribution:

Example 6: Let $x_1, ..., x_n$ be a random sample from $N(\theta, 100)$. Show that $W = \{(x_1, ..., x_n) : \bar{x} \geq C\}$ is a BCR for testing $H_0 : \theta = 75$ against $H_1 : \theta = 78$. Find n and C so that the size of the test $\alpha = 0.05$ and the power of the test $(1-\beta) = 0.90$.

Solution: The p.d.f. of $N(\theta, 100)$ is

$$f(x, \theta) = \left\{ \frac{1}{2\pi(100)} \right\}^{\frac{1}{2}} e^{-\frac{1}{2(100)}(x-\theta)^2} \quad ; -\infty < x < \infty$$

The likelihood function L is

$$L = \left(\frac{1}{200\pi} \right)^{n/2} e^{-\frac{1}{200}\sum_{i=1}^{n}(x_i-\theta)^2}$$

$$L_0 = \left(\frac{1}{200\pi} \right)^{n/2} e^{-\frac{1}{200}\sum(x_i-75)^2} \quad \text{under } H_0$$

$$L_1 = \left(\frac{1}{200\pi} \right)^{n/2} e^{-\frac{1}{200}\sum(x_i-78)^2} \quad \text{under } H_1$$

Using N - P lemma the BCR is obtained by considering $\frac{L_1}{L_0} \geq K$, where K is a positive constant.

$$\frac{L_1}{L_0} \geq K$$

$$\Rightarrow \frac{(200\pi)^{-n/2}\; e^{-\frac{1}{200}\sum(x_i-78)^2}}{(200\pi)^{-n/2}\; e^{-\frac{1}{200}\sum(x_i-75)^2}} \geq K$$

$$\Rightarrow e^{-\frac{1}{200}\left[\sum(x_i-78)^2-\sum(x_i-75)^2\right]} \geq K$$

Taking 'log' on both sides

$$-\frac{1}{200}\left(459n - 6\sum x_i\right) \geq \log K$$

$$\Rightarrow 6\sum x_i - 459\, n \geq 200 \log K$$

$$\Rightarrow \sum x_i \geq \frac{200 \log K + 459n}{6}$$

$$\Rightarrow \bar{x} = \frac{\sum x_i}{n} \geq \frac{200 \log K + 459\, n}{6n}$$

$\therefore$ The BCR for testing $H_0 : \theta = 75$ against $H_1 : \theta = 78$ is given by

$W = \left\{(x_1,..........., x_n) : \bar{x} \geq C\right\}$, where $C = \dfrac{200 \log K + 459\, n}{6n}$.

We are given

$$P\left(\bar{x} \geq C \;\middle|\; H_0\right) = \alpha = 0.05 \qquad .... (1)$$

and

$$P\left(\bar{x} \geq C \;\middle|\; H_1\right) = 1-\beta = 0.90 \qquad .... (2)$$

$X \sim N(\theta, 100)$ so that $\sigma^2 = 100, \sigma = 10$

$$\therefore \bar{x} \sim N\left(\theta, \frac{100}{n}\right).$$

Now, equation (1) can be written as

$$P\left[\frac{\bar{x}-\theta}{\sigma/\sqrt{n}} \geq \frac{C-75}{10/\sqrt{n}} \middle| H_0\right] = 0.05$$

$$\Rightarrow P\left[Z \geq \frac{C-75}{10/\sqrt{n}}\right] = 0.05$$, where Z is the standard normal variate.

From standard normal distribution tables we have

$$\frac{C-75}{10/\sqrt{n}} = 1.645 \qquad \text{.......................} (3)$$

Equation (2) can be written as

$$P\left[\frac{\bar{x}-\theta}{\sigma/\sqrt{n}} \geq \frac{C-78}{10/\sqrt{n}} \middle| H_1\right] = 0.90$$

From standard normal distribution tables we have

$$\frac{C-78}{10/\sqrt{n}} = -1.282 \qquad \text{.......................} (4)$$

Solving equations (3) and (4) we have

$n = 95.19$ and $C = 76.7$

Example 7: Derive the best critical region for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$ of the normal population mean μ when the variance σ^2 is known, based on a sample of size n drawn from $N(\mu, \sigma^2)$.

Solution: Let $x_1, ..., x_n$ be a random sample drawn from a normal population with p.d.f. is

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

The likelihood function L is

$$L = \underset{i=1}{\overset{n}{\pi}} f\left(x_i,\ \mu, \sigma^2\right)$$

$$L = \left(\frac{1}{\sqrt{2\pi}\ \sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$L_0 = \left(\frac{1}{\sqrt{2\pi}\ \sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu_0)^2} \quad \text{under } H_0$$

$$L_1 = \left(\frac{1}{\sqrt{2\pi}\ \sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu_1)^2} \quad \text{under } H_1$$

Using N - P lemma the BCR is obtained by considering $\frac{L_1}{L_0} \geq K$, where K is a positive constant.

Now, $\frac{L_1}{L_0} \geq K$

$$\frac{\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu_1)^2}}{\left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu_0)^2}} \geq K$$

$$\Rightarrow \quad e^{-\frac{1}{2\sigma^2}\left\{\sum(x_i-\mu_1)^2 - \sum(x_i-\mu_0)^2\right\}} \geq K$$

Taking 'log' on both sides

$$-\frac{1}{2\sigma^2}\left\{\sum(x_i-\mu_1)^2 - \sum(x_i-\mu_0)^2\right\} \geq \log K$$

$$-\left\{\left(\sum x_i^2 + \sum\mu_1^2 - 2\mu_1\sum x_i\right) - \left(\sum x_i^2 + \sum\mu_0^2 - 2\mu_0\sum x_i\right)\right\} \geq 2\sigma^2 \log K$$

$$\Rightarrow 2\sum x_i\,(\mu_1-\mu_0) + n\left(\mu_0^2-\mu_1^2\right) \geq 2\sigma^2 \log K$$

$$\Rightarrow 2\sum x_i\,(\mu_1-\mu_0) \geq 2\sigma^2 \log K + n\left(\mu_1^2-\mu_0^2\right)$$

$$\Rightarrow \sum x_i \left(\mu_1 - \mu_0\right) \geq \frac{2\sigma^2 \log K + n\left(\mu_1^2 - \mu_0^2\right)}{2}$$

If $\mu_1 > \mu_0$

$$\sum x_i \geq \frac{\sigma^2 \log K}{\left(\mu_1 - \mu_0\right)} + \frac{n\left(\mu_1 + \mu_0\right)}{2}$$

$$\Rightarrow \bar{x} = \frac{\sum x_i}{n} \geq \frac{\sigma^2 \log K}{n\left(\mu_1 - \mu_0\right)} + \frac{\left(\mu_1 + \mu_0\right)}{2}$$

$\therefore$ The BCR when $\mu_1 > \mu_0$ is given by

$$W = \left\{\left(x_1, \ldots\ldots\ldots\ldots, x_n\right) : \bar{x} \geq \frac{\sigma^2 \log K}{n\left(\mu_1 - \mu_0\right)} + \frac{\left(\mu_1 + \mu_0\right)}{2}\right\}$$

If $\mu_1 < \mu_0$, the BCR is given by

$$W = \left\{\left(x_1, \ldots\ldots\ldots\ldots\ldots, x_n\right) : \bar{x} \leq \frac{\sigma^2 \log K}{n\left(\mu_1 - \mu_0\right)} + \frac{\left(\mu_1 + \mu_0\right)}{2}\right\}$$

Example 8: Let $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ be a random sample from $N\left(\mu, \sigma^2\right)$. Obtain the best critical region for testing $H_0 : \mu = 0,\ \sigma^2 = 1$ against $H_1 : \mu = 1,\ \sigma^2 = 4$.

Solution: The p.d.f. of normal distribution is

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

The likelihood function L is

$$L = \left(\frac{1}{\sqrt{2\pi}\ \sigma}\right)^n\ e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2}$$

$$L_0 = \left(\frac{1}{\sqrt{2\pi}}\right)^n\ e^{-\frac{1}{2}\sum x_i^2} \qquad \text{under } H_0$$

$$L_1 = \left(\frac{1}{2\sqrt{2\pi}}\right)^n e^{-\frac{1}{2(4)}\sum(x_i-1)^2} \quad \text{under } H_1$$

Using N - P lemma the BCR is obtained by considering $\frac{L_0}{L_1} \le K$, where K is a positive constant.

So $\frac{L_0}{L_1} \le K$

$$\Rightarrow \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum x_i^2}}{\frac{1}{2^n\left(\sqrt{2\pi}\right)^n} e^{-\frac{1}{8}\sum(x_i-1)^2}} \le K$$

$$\Rightarrow 2^n \; e^{-\frac{1}{2}\sum x_i^2 + \frac{1}{8}\sum(x_i-1)^2} \le K$$

Taking 'log' on both sides

$$n \log 2 - \frac{1}{2}\sum x_i^2 + \frac{1}{8}\sum\left(x_i - 1\right)^2 \le \log K$$

$$-\frac{1}{2}\sum x_i^2 + \frac{1}{8}\sum x_i^2 + \frac{n}{8} - \frac{2\sum x_i}{8} \le \log K - n \log 2$$

$$-\frac{3}{8}\sum x_i^2 - \frac{1}{4}\sum x_i \le \log K - n \log 2 - \frac{n}{8}$$

$$\frac{3}{8}\sum x_i^2 + \frac{1}{4}\sum x_i \ge \frac{n}{8} + n \log 2 - \log K$$

$$\Rightarrow 3\sum x_i^2 + 2\sum x_i \ge n + 8n \log 2 - 8 \log K$$

$\therefore$ The BCR for testing $H_0 : \mu = 0, \sigma^2 = 1$ against $H_1 : \mu = 1, \sigma^2 = 4$ is

$$W = \left\{\left(x_1, \ldots\ldots\ldots\ldots, x_n\right) : 2\sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i \ge C\right\}$$

where $C = n + 8n \log 2 - 8 \log K$.

10.6 Exercises:

1. Suppose you are testing $H_0 : \lambda = 2$ against $H_1 : \lambda = 1$, where λ is the parameter of the Poisson distribution. Obtain the best critical region of the test using N - P lemma.

2. Let X have a p.d.f. of the form

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}, \quad 0 < x < \infty,\ \theta > 0$$

Obtain the BCR for testing $H_0 : \theta = 2$ against $H_1 : \theta = 1$, using a random sample of 2 observations.

3. Obtain the BCR for testing the mean $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1\ (\mu_1 > \mu_0)$ when $\sigma^2 = 1$ in a normal population based on a random sample of size n drawn from it.

4. Obtain the best critical region for testing $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{1}{4}$ where p is the parameter in binomial distribution, using a sample size n = 3.

5. On the basis of a single observation from the population which follows exponential distribution, we have to test $H_0 : \theta = 1$ against $H_1 : \theta = 4$ and we reject H_0 when $x > 3$. Find the size and power of the test.

6. Find the best critical region for testing $H_0 : \theta = 0$ against $H_1 : \theta = 4$ when $\sigma^2 = 1$, based on a random sample of size n drown from $N(\theta, \sigma^2)$.

10.7 Summary:

Neyman - Pearson lemma is used to derive best critical regions for testing a simple hypothesis against a simple alternative hypothesis for different parameters in binomial distribution, Poisson distribution, exponential distribution and normal distribution (with known S.D.). Some exercises are given to the students to try on their own.

10.8 Technical Terms:

Likelihood function

Best critical region

Lesson Writer

C.V. RAO

Lesson 11

LARGE SAMPLE TESTS - 1

Objective:

After studying this lesson the student will be conversant with the concepts of central limit theorem, critical values in one - tailed and two - tailed tests, procedure in testing of hypothesis in general and in particular test of significance for single mean and test of significance for difference of two means.

Structure of the lesson:

11.1 Introduction:

Once the sample data has been gathered through an observational study, statistical inference allows analysts to assess some claim about the population from which the sample has been drawn. The methods of inference used to support or reject the claims based on sample data are known as tests of significance.

If the population is normal and if we want to test any parameter of the normal population hypothetically, we can use both large sample tests and small sample tests. The selection of these tests depend on the size of the sample. If the size of the sample is greater than or equal to 30, we treat it as a large sample and a test procedure based on large sample is called large sample test. Here we will study large sample tests for testing a single mean and difference of two means.

11.2 Central Limit Theorem:

If $X_1, X_2, \cdots\cdots\cdots\cdots, X_n$ be independent random variables such that $E(X_i) = \mu_1$ and $V(X_i) = \sigma_i^2$, then under certain very general conditions, the random variable $S_n = X_1 + X_2 + \ldots\ldots\ldots + X_n$ is asymptotically normal with mean μ and standard deviation σ where $\mu = \sum_{i=1}^{n} \mu_i$ and $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$. In other words, $Z = \frac{S_n - \mu}{\sigma}$ approaches $N(0,1)$ as $n \to \infty$.

If 't' is any statistic used to test null hypothesis against alternative hypothesis and is computed from a large sample drawn from the population. Then the sampling distribution of 't' is asymptotically normally distributed.

i.e., $t \sim N\left[E(t),\ V(t)\right]$ as $n \to \infty \cdot$

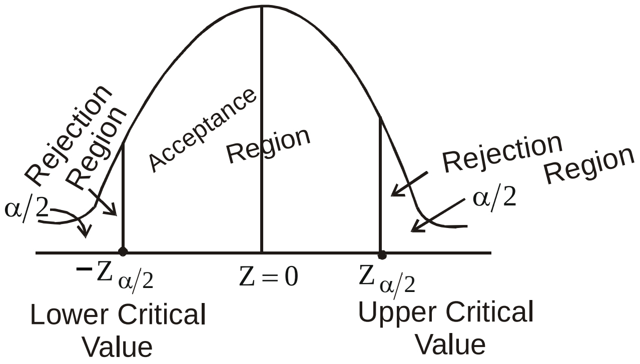
The standard normal variate

$$Z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0,1).$$

The central limit theorem implies that the sum of n independently distributed random variables is approximately normal, whatever the distribution of individual variables may be. i.e., either normal or non-normal. This theorem is quite useful in solving the problems of testing of hypothesis and in interval estimation of parameters.

11.3 Critical Value or Significant Value:

If α is the size of the critical region, then the value of the test statistic Z at α denoted by Z_α at which the sample space is divided into two disjoint regions called acceptance and rejection regions, then Z_α is called a critical value and is based upon the size 'α' and the alternative hypothesis H_1, whether it is one - tailed or two - tailed test.



For two - tailed test - if the size of the critical region is α, then half of the critical region $(\alpha/2)$ would be to the left side and the other half $(\alpha/2)$ to the right side of the probability curve.

i.e., $\Pr\left\{Z \geq Z_{\alpha/2}\right\} = \alpha/2$

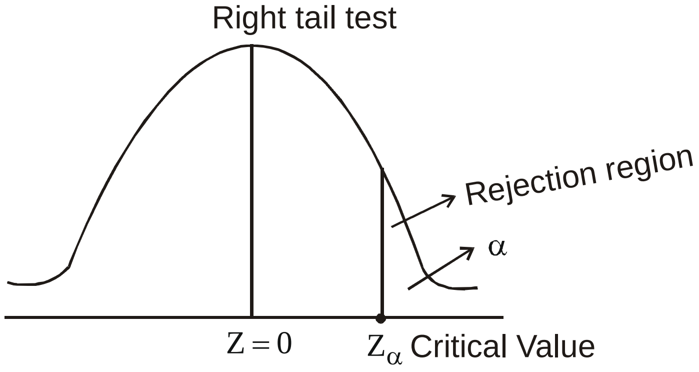
and $\Pr\left\{Z \leq -Z_{\alpha/2}\right\} = \alpha/2$

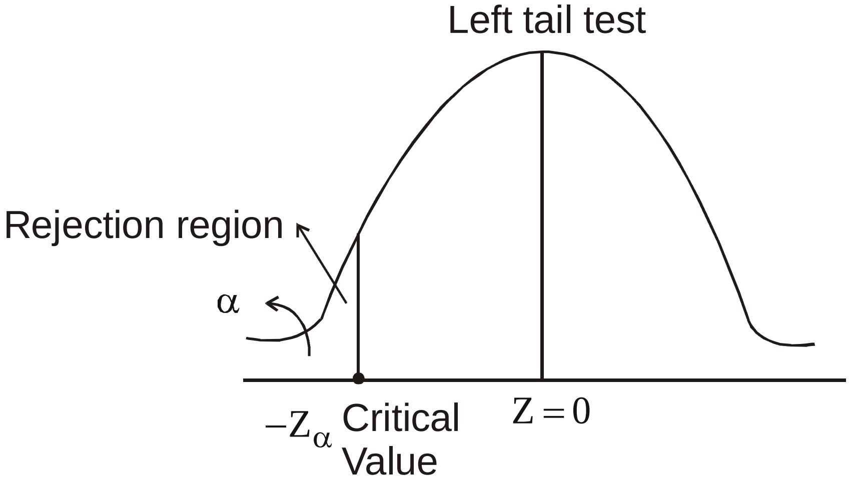
$\Rightarrow \Pr\left\{Z \geq Z_{\alpha/2}\right\} + P_r\left\{Z \leq -Z_{\alpha/2}\right\} = \alpha$

$\Rightarrow \Pr\left\{|Z| \geq Z_{\alpha/2}\right\} = \alpha$ (Rejection region)

$\Rightarrow \Pr\left\{|Z| < Z_{\alpha/2}\right\} = 1-\alpha$ (Acceptance region)

For one - tail test i.e., either right tail or left tail test, the size of the critical region 'α' will be either in the right tail or in left tail only.





For right tail test $\Pr\{Z \geq Z_\alpha\} = \alpha$

and $\Pr\{Z < Z_\alpha\} = 1-\alpha$

and for left tail test

$\Pr\{Z \leq -Z_\alpha\} = \alpha$

and $\Pr\{Z > Z_\alpha\} = 1-\alpha$.

The critical value for two tailed test with l.o.s. 2α is equal to the critical value of one tail test at the l.o.s. α.

The critical values for $\alpha = 1\%$ and $\alpha = 5\%$ are given as

Values of Z_α

Type of test	$\alpha = 1\%$	$\alpha = 5\%$
One - tailed	2.33	1.645
Two - tailed	2.58	1.96

11.4 Tests of Significance :

Tests of significance based on sample observations enable us to test

(1) The deviation between the observed sample statistic and the hypothetical parameter value is significant or not,

(2) The deviation between two sample statistics is significant or not.

11.5 Procedure for testing of hypothesis:

The following are the important steps to be followed in any problem of testing of hypothesis.

(1) Set up the null hypothesis

(2) Setup the alternative hypothesis which enables us to decide whether we have to use a test of one - tail or two tail.

(3) Choose an appropriate level of significance α and select the critical value Z_α. α is to be choosen before the sample is drawn.

(4) Choose an appropriate statistic 't' and compute the standard normal test criteria

$$Z = \frac{t - E(t)}{\sqrt{V(t)}}$$ under H_0.

(5) **Conclusion:** If the calculated value of $|Z|$ is less than the critical value of Z i.e., Z_α, we accept H_0, otherwise we reject H_0.

11.6 Sampling of Variables:

In any population if we want to study a character which can be measured quantitatively, such as height, age, income, etc., we apply sampling of variables by defining the measurable character as a variable. Since, the variable can be represented numerically, we have populations of numerical data for which we can test the statistical measures like mean, S.D., etc.

The following gives the detailed explanation of test of single mean and difference of means.

11.7 Test of Single Mean:

Let $X_1, X_2, \ldots\ldots\ldots, X_n$ be a large sample drawn at random from a population with mean μ and variance σ^2. Let $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ be the sample mean. Then the sampling distribution of $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$ for large samples.

Suppose we want to test the population mean $\mu = \mu_0$ (or) test the significant difference between sample mean and population mean, we set up the statistical hypothesis as follows:

H_0 : There is no significant difference between the sample mean and the population mean

i.e., $H_0 : \mu = \mu_0$

Vs

H_1 : There is a significant difference between the sample mean and the population mean

i.e., $H_1 : \mu \neq \mu_0$

The required test statistic to test the above hypothesis H_0 is

$$Z = \frac{\bar{x} - E(\bar{x})}{S.E.(\bar{x})} \sim N(0,1)$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ under } H_0.$$

When σ is not known, replace it by $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$

$$\Rightarrow Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim N(0,1) \text{ under } H_0.$$

Conclusion: If the calculated value of $|Z|$ is less than the critical value of Z at level of significance α , we accept H_0 and conclude that there is no significant difference between sample mean and population mean, otherwise reject H_0 and conclude that there is a significant difference between sample mean and population mean.

11.8 Confidence Limits for μ:

We know that

$$\Pr\left\{|Z| < Z_{\alpha/2}\right\} = 1-\alpha$$

$100(1-\alpha)\%$ confidence interval for μ is given by

$$\Pr\left\{|Z| < Z_{\alpha/2}\right\} = 1-\alpha$$

$$\Rightarrow \Pr\left\{\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| < Z_{\alpha/2}\right\} = 1-\alpha$$

$$\Rightarrow \Pr\left\{|\bar{x}-\mu| < \sigma/\sqrt{n} \cdot Z_{\alpha/2}\right\} = 1-\alpha$$

$$\Rightarrow \Pr\left\{\bar{x}-Z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \bar{x} + Z_{\alpha/2} \cdot \sigma/\sqrt{n}\right\} = 1-\alpha$$

$\left\{\bar{x} - Z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \bar{x} + Z_{\alpha/2} \cdot \sigma/\sqrt{n}\right\} = 1-\alpha$ is called $100(1-\alpha)\%$

confidence interval and $\bar{x} \pm Z_{\alpha/2} \cdot \sigma/\sqrt{n}$ are called the $100(1-\alpha)\%$ confidence limits.

11.9 95% confidence interval for μ:

For $\alpha = 0.05$, we have $Z_{\alpha/2} = 1.96$, then

$$\Pr\left\{\bar{x} - 1.96\left(\sigma/\sqrt{n}\right) < \mu < \bar{x} + 1.96\left(\sigma/\sqrt{n}\right)\right\} = 0.95$$

Therefore 95% confidence interval for μ is

$$\left(\bar{x} - 1.96\left(\sigma/\sqrt{n}\right) < \mu < \bar{x} + (1.96)\sigma/\sqrt{n}\right)$$

and 95% confidence limits for μ are

$$\bar{x} \pm 1.96\left(\sigma/\sqrt{n}\right).$$

11.10 99% confidence interval for μ:

For $\alpha = 0.01,\ Z_{\alpha/2} = 2.58$, then

$$\Pr\left\{\bar{x} - (2.58)\sigma/\sqrt{n} < \mu < \bar{x} + (2.58)\,\sigma/\sqrt{n}\right\} = 0.99$$

$\therefore$ 99% confidence interval for μ is

$$\left(\bar{x} - (2.58)\sigma/\sqrt{n} < \mu < \bar{x} + (2.58)\sigma/\sqrt{n}\right)$$

and 99% confidence limits for μ are

$$\left(\bar{x} \pm 2.58\ \sigma/\sqrt{n}\right).$$

Problem 1: A sample of 900 members has a mean 3.4 cm and S.D. 2.61 cm, drawn from a population of mean 3.25 cm and S.D. 2.61 cm. Test the significance of sample mean at 5% l.o.s. Also, find 95% confidence limits for true mean.

Solution: The hypothesis to be tested for the above problem is

H_0 : The sample has been drawn from the population with mean $\mu = 3.25$.

H_1 : The sample has not been drawn from the population.

i.e., H_1 : $\mu \neq 3.25$ (Two - tailed)

Given $\bar{x} = 3.4$ cm

$s = 2.61$ cm ; $n = 900$

$\mu = 3.25$ cm

$\sigma = 2.61$ cm

The required test statistic to test H_0 is

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{3.40 - 3.25}{2.61/\sqrt{900}}$$

$$\Rightarrow Z = 1.73$$

Tabulated value of Z for two tailed test at 5% l.o.s. is 1.96

Conclusion: Since, the calculated value of Z is less than the tabulated value of Z at 5% l.o.s. we accept H_0 and conclude that the sample is drawn from the population with mean $\mu = 3.25$ cm.

95% confidence limits for μ are given by

$$\bar{x} \pm 1.96\left(\sigma/\sqrt{n}\right)$$

$$\Rightarrow 3.40 \pm 1.96\left(\frac{2.61}{\sqrt{900}}\right) = 3.40 \pm 0.1705$$

i.e., 3.2295 and 3.5705 are the 95% confidence limits for μ to lie.

Problem 2: An ambulance service claims that it takes on the average 8.9 minutes to reach its destination in emergency calls. To check on this claim, the agency which licences ambulance service has done survey on 50 emergency calls with a mean of 9.3 minutes and S.D. of 1.6 minutes. What is your conclusion at 1% l.o.s.

Solution: The hypothesis to be tested for the above problem is

H_0 : The average time to reach the destination is 8.9 minutes. i.e., $H_0 : \mu = 8.9$

Vs

H_1 : The average time to reach the destination is not 8.9 minutes.

i.e., $H_1 : \mu \neq 8.9$

Also, we have

$$n = 50 \; ; \mu = 8.9 \; ; \bar{x} = 9.3 \quad \& \quad s = 1.6$$

The required test statistic to test the claim is

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim N(0,1) \text{ at } \alpha\% \text{ l.o.s.}$$

$$Z = \frac{9.3 - 8.9}{(1.6)/\sqrt{50}} = 1.767$$

Tabulated value of Z at 1% l.o.s. is 2.58

Conclusion: Since, the calculated value of Z is less than the tabulated value of Z at 1% l.o.s., we accept H_0 and conclude that the average time taken by the ambulance service to reach the destination on emergency call is 8.9 minutes.

Problem 3: A normal population has mean of 0.1 and standard deviation of 2.1. Find the probability that mean of a sample of size 900 will be negative.

Solution: Here we are given that $X \sim N\left(\mu, \sigma^2\right)$ with mean $\mu = 0.1$ and $\sigma = 2.1$. Also $n = 900$.

Since, $X \sim N\left(\mu, \sigma^2\right)$ the sample mean $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$.

The standard normal variate corresponding to $\bar{x}$ is given by

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0.1}{2.1/\sqrt{900}} = \frac{\bar{x} - 0.1}{0.07}, \text{ where } Z \sim N(0,1)$$

$$\Rightarrow \bar{x} = 0.1 + (0.07)Z$$

The required probability that the mean of a sample of size 900 will be negative is

$$\Pr\left(\bar{x} < 0\right) = \Pr\left(0.1 + (0.07)Z < 0\right)$$

$$= \Pr\left(Z < \frac{-0.1}{0.07}\right)$$

$$= \Pr\left(Z < -1.43\right)$$

$$= \Pr\left(Z \geq 1.43\right)$$

$$= 0.5 - \Pr\left\{0 < Z < 1.43\right\}$$

$$= 0.5 - 0.4236 \qquad \text{(From normal distribution tables)}$$

$$= 0.0764$$

$\therefore$ Probability that mean of a sample of size 900 will be negative is 0.0764.

Problem 4: The guaranteed average life of a certain type of electric light bulbs is 1000 hours with a standard deviation of 125 hours. It is decided to sample the output so as to ensure that 90 percent of bulbs do not fall short of the guaranteed average by more than 2.5 percent. What must the minimum size of the sample?

Solution: Here we are given the average life of certain bulbs as 1000 hours.

i.e., $\mu = 1000$

and the standard deviation as $\sigma = 125$ hours.

Since we do not want the sample mean to be less than the guaranteed average mean $(\mu = 1000)$ by more than 2.5% we should have

$$\bar{x} > 1000 - (2 \cdot 5 \times 1000/100)$$

$$\bar{x} > 975$$

Let n be the given sample size. Then

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ for large sample.}$$

We want $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} > \frac{975 - 1000}{125/\sqrt{n}} = \frac{-\sqrt{n}}{5}$ $\qquad (\because \bar{x} > 975)$

According to the given condition $P\left(Z > -\frac{\sqrt{n}}{5}\right) = 0.90$

$$\Rightarrow P\left(-\frac{\sqrt{n}}{5} < Z < 0\right) + P(0 < Z < \alpha) = 0.90$$

$$\Rightarrow P\left(0 < Z < \frac{\sqrt{n}}{5}\right) + 0.5 = 0.90$$

$$\Rightarrow P\left(0 < Z < \frac{\sqrt{n}}{5}\right) = 0.40$$

$$\Rightarrow \frac{\sqrt{n}}{5} = 1.28 \Rightarrow n = 25 \times (1.28)^2 = 40.96 \simeq 41.$$

Therefore, the required minimum sample size to satisfy the given condition is 41.

11.11 Test of difference of two means:

Let $\overline{x_1}$ be the sample mean of sample size n_1 observations drawn at random from a population with mean μ_1 and variance σ_1^2.

Let $\overline{x_2}$ be the sample mean of sample size n_2 observations drawn at random from another population with mean μ_2 and variance σ_2^2.

Then $\overline{x_1} \sim N\left(\mu_1 , \sigma_1^2/n_1\right) \; \forall$ large samples

$\overline{x_2} \sim N\left(\mu_2 , \sigma_2^2/n_2\right) \; \forall$ large samples

$$\therefore \; \overline{x_1} - \overline{x_2} \sim N\left(\mu_1 - \mu_2 , \; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Suppose we want to test whether the two population means μ_1 and μ_2 are equal or to test whether there is any significant difference between two sample menas $\overline{x_1}$ and $\bar{x}_2$, the required statistical hypothesis to test is

H_0 : The two population means are equal

i.e., $H_0 \; : \mu_1 = \mu_2$

Vs

H_1 : Two population means are not equal

i.e., $H_1 \; : \mu_1 \neq \mu_2$.

The required standard normal test statistic to test the above hypothesis is

$$Z = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - E\left(\bar{x}_1 - \bar{x}_2\right)}{S.E.\left(\bar{x}_1 - \bar{x}_2\right)} \sim N(0,1)$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Under H_0, $\mu_1 = \mu_2$, then

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \cdot$$

If σ_1^2 and σ_2^2 are not known, then they are replaced by the first sample variance S_1^2 and the second sample S_2^2 respectively and now

$$Z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1) \text{ under } H_0 .$$

If the two population variances are equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Then $Z = \dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$ under H_0

If two population variances are equal and not known

i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown)

Then σ^2 is replaced by its unbiased estimator of pooled sample variances i.e., $S^2 = \dfrac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}$.

If the calculated value of $|Z|$ is less than the tabulated value of Z at $\alpha\%$ level of significance we accept H_0 and conclude that there is no significant difference between two population means, otherwise reject H_0.

Problem 1: The mean heights of two large samples of 1000 and 2000 members are 67.5 inches and 68 inches respectively. Can the samples be regarded as drawn from the same normal population with standard deviation of 2.5 inches.

Solution: The hypothesis to be tested for the problem is

H_0 : The two samples have come from the same population

Vs

H_1 : The two samples have not come from the same population.

Here, we are given

$n_1 = 1000$; $n_2 = 2000$

$\overline{x_1} = 67.5$; $\overline{x_2} = 68$

& $\sigma = 2.5$

The required test statistic to test the above hypothesis is

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

$$Z = \frac{67.5 - 68}{2.5\sqrt{\frac{1}{1000} + \frac{1}{2000}}} = \frac{-0.5}{2.5(0.0387)} = -5.165$$

$|Z| = 5.165$. The table value of Z at 5% l.o.s. is 1.96

Since, the calculated value of Z is greater than the tabulated value of Z at 5% l.o.s. we reject H_0 and conclude that the two samples have not come from the same population.

Problem 2: In a survey of buying habits, 400 women shoppers were choosen at random in a supermarket 'A' located in a certain section of city, their average weekly expeniture on food is Rs 250/- with a S.D. of Rs 40/-. For another 400 women shoppers choosen at random in a supermarket 'B' in another section of city, the average weekly food expenditure is Rs 220/- with a S.D. of Rs 55/-. Test at 1% l.o.s. whether average weekly food expenditure of the women in supermarket A is more than the weekly average food expenditure of the women in supermarket B.

Solution: Let μ_1 represent the average food expenditure of women from supermarket A.

Let μ_2 represent the average food expenditure of women of supermarket B.

The hypothesis to be tested for the above problem is

H_0 : There is no significant difference in the average expenditure on food of women of two supermarkets.

i.e., $H_0 : \mu_1 = \mu_2$

H_1 : The average food expenditure of women of supermarket A is more than the average food expenditure of women of supermarket B.

i.e., $H_1 : \mu_1 > \mu_2$

Also, we are given

$\overline{x_1} = 250$; $\overline{x_2} = 220$

$n_1 = 400$; $n_2 = 400$

$S_1 = 40$; $S_2 = 55$

The required test statistic to test the above hypothesis is

$$Z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1) \text{ under } H_0$$

$$Z = \frac{250 - 220}{\sqrt{\frac{1600}{400} + \frac{3025}{400}}} = \frac{30}{3.4} = 8.823$$

The tabulated value of Z at 1% l.o.s. for one - tail test is 2.33

Since, the calculated value of Z is greater than the tabulated value of Z at 1% level of significance, we reject H_0 and conclude that the average food expenditure of women of supermarket A is more than the average food expenditure of women of supermarket B.

Problem 3: A store keeper wants a large quantity of bulbs from two brands labelled I and II. He bought 100 bulbs of each brand and found by testing that brand I have mean life time of 1120 hours and a S.D. of 75 hours, brand II had mean life time of 1062 hours and S.D. of 82 hours. Assuming that the lots of brand I and brand II have the same S.D. examine whether the quality of brand II is inferior to the quality of brand I.

Solution: Let μ_1 represents the average life time of brand I bulbs, μ_2 represents the average life time of brand II bulbs. The hypothesis to be tested for the given problem is

H_0 : The average life time of both the brands of bulbs are equal

i.e., $H_0 : \mu_1 = \mu_2$

Vs

H_1 : The average life time of brand I bulbs is greater than that of bulbs of brand II.

i.e., $H_1 : \mu_1 > \mu_2$

Given

$$\bar{x}_1 = 1120 \quad ; \quad \bar{x}_2 = 1062$$

$$n_1 = 100 \quad ; \quad n_2 = 100$$

$$S_1 = 75 \quad ; \quad S_2 = 82$$

The test statistic to test the above hypothesis with equal S.D.'s but unknown value is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}$$

$$S^2 = \frac{100(75)^2 + 100(82)^2}{100 + 100} = 6174.5$$

$$Z = \frac{1120 - 1062}{\sqrt{6174.5\left(\frac{1}{100} + \frac{1}{100}\right)}} = \frac{58}{5.556 \times 2} = 5.2193$$

The tabulated value of Z for one tail test at 5% l.o.s. is 1.645.

Since the calculated Z > tabulated Z at 5% l.o.s., we reject H_0 and conclude that quality of brand II bulbs is inferior to the quality of brand I bulbs.

Problem 4: A random sample of 1200 men with their mean pay of Rs 400/- is taken from a state with an S.D. = Rs 60/- Another random sample of 1000 men with their mean pay of Rs 500/- is taken from another state with an S.D. of Rs 80/-. Discuss whether the mean levels of two states are different.

Solution: Let μ_1 represent the mean pay of first state and μ_2 represent the mean pay of the second state. The hypothesis to be tested is

H_0 : There is no significant difference between mean pays of two states.

i.e., $H_0 : \mu_1 = \mu_2$

H_1 : There is a significant difference between mean pays of two states.

i.e., $H_1 : \mu_1 \neq \mu_2$

Given

$n_1 = 1200$; $n_2 = 1000$

$\bar{x}_1 = 400$; $\bar{x}_2 = 500$

$\sigma_1 = 60$; $\sigma_2 = 80$

The test statistic to test the above hypothesis is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$Z = \frac{400 - 500}{\sqrt{\frac{(60)^2}{1200} + \frac{(80)^2}{1000}}} \sim N(0,1)$$

$$Z = \frac{-100}{\sqrt{3 + 6.4}} = \frac{-100}{3.0659} = -32.6168$$

$$|Z| = 32.6168$$

Tabulated value of Z at 5% l.o.s. for two - tailed test is 1.96

Since the calculated value of Z is greater than the tabulated value of Z at 5% l.o.s., we reject H_0 and conclude that there is a significant difference in the mean pay of men of two states.

11.12 Exercises:

1. A sample of 100 students is taken from a large population. The mean height of these students is 64 inches and the standard deviation is 4 inches. Can it reasonably be regarded that in the population, mean height is 66 inches.

2. A sample of 400 individuals is found to have a mean height of 67.47 inches. Can it be reasonbly regarded as a sample from a large population with mean height of 67.39 inches and standard deviation 1.30 inches.

3. The mean breaking strength of cables supplied by the manufacturer is 1800, with a standard deviation 100. By a new technique in the manufacturing process it is claimed that the breaking strenth of the cables has increased. In order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strenth is 1850. Can we support the claim at 0.01 level of significance.

4. A sample of 450 items is taken from a population whose standard deviation is 20. The mean of the sample is 30. Test whether the sample has come from the population with mean 29. Also calculate 95% & 99% confidence limits of the population mean.

5. A sample of heights of 6,400 English men has a mean of 67.85 inches and S.D. of 2.56 inches, while a sample of heights of 1600 Australians has a mean of 68.55 and a S.D. of 2.52 inches. Do the data indicate that Australians are on the average taller than Englishmen.

6. The mean yield of two sets of plots and their variability are as given below. Examine whether the difference in the mean yields of the two sets of plots is significant.

	Set of 40 plots	**Set of 60 plots**
Mean yield of plot	1258 lb	1243 lb
S.D. per plot	34	28

7. In a survey of incomes of two classes of workers, two random samples gave the following results. Examine whether the difference between means is significant.

	Size	**Mean annual income**	**S.D.**
Sample I	100	582	24
Sample II	100	546	28

11.13 Summary:

This lesson discusses in detail the large sample tests for variables - test for single mean and test for difference of two means. Also, the concept of central limit theorem to large samples, identification of critical region and critical values in one tailed and two tailed tests and the stepwise procedure for testing of hypothesis are explained. A number of problems are solved and a number of exercises are given to students to solve on their own.

11.14 Technical Terms:

Null & alternative hypothesis

Critical value

One tailed test

Two tailed test

Tests of significance

Sampling for variables

Test for single mean

Test for difference of two means

Lesson Writter

C.V.RAO

Lesson 12

LARGE SAMPLE TESTS - 2

Objective:

After studying this lesson the student will be conversant in the theory and applications of testing a single standard deviation, the difference of two standard deviations, a single proportion, the difference of two proportions.

Structure of the lesson:

12.1 Introduction:

In the previous lesson we have discussed the concept of sampling for variables with respect to the test for single mean and the difference of two means. In continuation to that now we will discuss the test for single S.D. and difference of two S.D.'s. Also the concept of sampling of attributes w.r.t. the test for single proportion and the difference of two proportions are discussed.

12.2 Test for single standard deviation:

Let $X_1, X_2, ..., X_n$ be a random sample drawn from the population. Let $s = \sqrt{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2}$ be the sample standard deviation, then the sampling distribution of S.D. is approximately normally distributed with mean σ and variance $\sigma^2/2n$ for large samples, where σ is the population S.D. i.e., $E(s) = \sigma \;\&\; V(s) = \sigma^2/2n$.

$\Rightarrow\; s \sim N\left(\sigma,\; \sigma^2/2n\right)$ in large samples.

Suppose we want to test the population standard deviation $\sigma = \sigma_0$ (say) or to test the significant difference between the sample S.D. and the population S.D., we set up the statistical hypothesis as

H_0 : There is no significant difference between the population S.D. and the sample S.D.

i.e., $H_0 : \sigma = \sigma_0$

Vs

H_1 : There is a significant difference between the population S.D. and the sample S.D.

i.e., $H_1 : \sigma \neq \sigma_0$

The required normal test criteria to test the above hypothesis is to compute

$$Z = \frac{s - E(s)}{S.E.(s)} \sim N(0,1) \qquad \text{under } H_0$$

$$Z = \frac{s - \sigma}{\sigma / \sqrt{2n}} \sim N(0,1) \qquad \text{under } H_0$$

$$Z = \frac{s - \sigma_0}{\sigma_0 / \sqrt{2n}} \sim N(0,1) \qquad \text{under } H_0$$

If the calculated $|Z|$ is less than the tabulated value of Z at α % l.o.s., we accept H_0 and conclude that there is no significant difference between the sample S.D. and the population S.D. otherwise reject H_0.

Problem 1: A sample of 2000 electric bulbs is found to have S.D. of 100 hours. Test the hypothesis that the S.D. of the population is 104 hours.

Solution: Let σ be the population S.D., the hypothesis to be tested for the above problem is

$H_0 : \sigma = 104$

Vs

$H_1 : \sigma \neq 104$.

Also, we are given $n = 2000$; $s = 100$

The required test criteria to test the above hypothesis is

$$Z = \frac{s-\sigma_0}{\sigma_0/\sqrt{2n}} \sim N(0,1) \qquad \text{under } H_0$$

$$Z = \frac{100-104}{104/\sqrt{4000}} \sim N(0,1)$$

$$Z = \frac{-4}{1.644} = -2.43$$

$$|Z| = 2.43$$

Tabulated value of Z for two tailed test at 5% l.o.s. is 1.96.

Since the calculated $|Z|$ is greater than the tabulated Z at 5% l.o.s. we reject H_0, and conclude that the population S.D. is not equal to 104 hours.

Problem 2: An instrument measured 30 units and found the S.D. of 30 measurements is 2.5. Test whether the efficiency of the instrument has increased in view of its general S.D. of measurement is 3.

Solution: Let σ be the population S.D. The hypothesis to be tested for the above problem is

$$H_0 \quad : \quad \sigma = 3$$

Vs

$$H_1 \quad : \quad \sigma < 3.$$

Also, we are given $n = 30, s = 2.5$

The required test criteria to test the above hypothesis is

$$Z = \frac{s-\sigma}{\sqrt{\sigma^2/2n}} \sim N(0,1) \qquad \text{under } H_0$$

$$Z = \frac{2.5-3}{\sqrt{9/60}} = \frac{-0.5}{3/\sqrt{60}} = -1.291$$

Since $H_1 : \sigma < 3$, this is a one tail test and the critical region lies in the left tail. We reject H_0 if $Z < -Z_\alpha$, where $-Z_\alpha$ is the critical value. Hence, for $\alpha = \cdot 05$, $-Z_\alpha = -1.645$.

Since the calculated Z is greater than the tabulated value of Z at 5% l.o.s., we accept H_0 and conclude that the efficiency of the instrument is not increased.

12.3 Test of significance for the difference of standard deviations:

Let s_1 and s_2 be the standard deviations of two independent samples of sizes n_1 and n_2 taken from two populations with their respective standard deviations σ_1 and σ_2. If we are interested to test whether the two population standard deviations are equal or not, we set up the statistical hypothesis as

H_0 : There is no significant difference between the two sample S.D.'s.

i.e., $H_0 : \sigma_1 = \sigma_2$

Vs

H_1 : There is a significant difference between the two sample S.D.'s

i.e., $H_1 : \sigma_1 \neq \sigma_2$

Now, the test criteria to test above H_0 is to compute

$$Z = \frac{(s_1 - s_2) - E(s_1 - s_2)}{S.E.(s_1 - s_2)} \sim N(0,1) \text{ for large samples.}$$

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}} \sim N(0,1) \text{ under } H_0.$$

σ_1^2 and σ_1^2 are usually unknown and for large samples, we use their estimates s_1^2 and s_2^2.

$$\therefore\ Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} \sim N(0,1) \text{ under } H_0.$$

If the calculated value of Z is less than the tabulated value of Z at $\alpha\%$ l.o.s., we accept H_0 and conclude that there is no significant difference between the two sample standard deviations, otherwise reject H_0.

Problem 1: Random samples drawn from two countries gave the following data relating to the heights of adult males:

	Country A	**Country B**
Mean height (in inches)	67.42	67.25
Standard deviation Iin (inches)	2.58	2.50
Number of males	1,000	1,200

Is the difference between standard deviations siginificant?

Solution: The hypothesis to be tested for the problem is

H_0 : There is no significant difference between the two sample standard deviations.

i.e., $H_0 : \sigma_1 = \sigma_2$

Vs

H_1 : There is a significant difference between the two sample standard deviations.

i.e., $H_1 : \sigma_1 \neq \sigma_2$

Also, we are given

$s_1 = 2.58$; $s_2 = 2.50$

$n_1 = 1000$; $n_2 = 1200$

The required test criteria to test H_0 Vs H_1 is to compute

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} \sim N(0,1)$$

$$Z = \frac{2.58 - 2.50}{\sqrt{\frac{(2.58)^2}{2\times1000} + \frac{(2.50)^2}{2\times1200}}} \sim N(0,1)$$

$$Z = \frac{0.08}{0.0774} = 1.03$$

Tabulated value of Z at 5% l.o.s. is 1.96

Since, the caluculated value of Z is less than the tabulated value of Z at 5% l.o.s., we accept H_0 and conclude that there is no significant difference between the sample standard deviations.

Problem 2: The sum of squares of the deviations from their arithmetic mean is 444 by a random sample of size 36 observations and it is 500 for a random sample of size 40 observations drawn from two different populations. Test whether the first population will have more S.D. than the other.

Solution: The hypothesis to be tested is

H_0 : The two population S.D.'s are equal

i.e., $H_0 : \sigma_1 = \sigma_2$

Vs

H_1 : The first population S.D. is greater than the second population S.D.

i.e., $H_1 : \sigma_1 > \sigma_2$

Given

$$\sum_{i=1}^{36}\left(X_i - \overline{X}\right)^2 = 444 \quad ; \quad \sum_{j=1}^{40}\left(Y_j - \overline{Y}\right)^2 = 500$$

$$s_1 = \sqrt{\frac{\Sigma\left(X_i - \overline{X}\right)^2}{n_1}} \quad \text{and} \quad s_2 = \sqrt{\frac{\Sigma\left(Y_j - \overline{Y}\right)^2}{n_2}}$$

$$s_1 = \sqrt{\frac{444}{36}} \quad ; \quad s_2 = \sqrt{\frac{500}{40}}$$

$$s_1 = 3.511 \quad ; \quad s_2 = 3.535$$

The required test criteria to test H_0 is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} \sim Z(0,1)$$

$$Z = \frac{3.511 - 3.535}{\sqrt{\frac{(3.511)^2}{2\times 36} + \frac{(3.535)^2}{2\times 40}}} = -0.04126$$

Since $H_1 : \sigma_1 > \sigma_2$, this is a one - tail test and the critical region lies in the right tail. We reject H_0 if $Z > Z_\alpha$.

Tabulated value of Z at 5% l.o.s. for one - tailed test is $Z_\alpha = 1.645$.

Since the calculated value of Z is less than the tabulated value of Z at 5% l.o.s., we accept H_0 and conclude that the two population S.D.'s are equal.

12.4 Fisher's Z - transformation:

The sampling distribution of sample correlation coefficient 'r' does not follow normal distribution even for large samples, but approximately follows normal distribution in certain situations. As an alternative to this, Fisher suggested a transformation denoted by

$Z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = \tanh^{-1} r$ which is exactly normally distributed with mean $E(Z) = \frac{1}{2}\log_e\left[\frac{1+\rho}{1-\rho}\right] = \tanh^{-1}\rho$ and variance $V(Z) = \frac{1}{n-3}$ for large samples.

Particularly, this transformation is fairly good when $n > 50$, provided the population is bivariate normal population with $\rho \neq 0$ and $\rho \neq \pm 1$.

12.5 Applications of Z - transformation:

Z - transformation has following applications in statistics.

(1) To test if an observed value of $'r'$ differs significantly from a hypothetical value ρ of the population correlation coefficient.

(2) To test the significance of the difference between two independent sample correlation coefficients.

(3) To obtain pooled estimate of ρ.

12.6 Sampling of Attributes:

Consider a sample of size n drawn from a population which is divided into two complementary classes, one possessing a particular attribute and the other not possessing that attribute. Then note down the number of individuals possessing that attribute in the sample. The presence of the attribute in an individual is termed as success and its absence as failure. In this case, a sample of 'n' individuals is identified with that of a series of 'n' independent Bernoulli trails with constant probability of success P for each trail. Then the probability of x success out of 'n' independent Bernoulli trails is a Binomial probability given by

$$P(X=x) = {}^nC_x\, P^x\,(1-P)^{n-x} \;\forall\; x = 0,1,2,............,n$$

with $E(X) = nP\,;\; V(X) = nP(1-P)\cdot$

12.7 Test of Single Proportion:

Let $x_1, x_2, \dots\dots\dots, x_n$ be a random sample drawn from a population with proportion P of an attribute. Let X denote number of individuals possessing the attribute with a sample of n individuals. Then $X \sim B(n, p)$.

Let $p = \frac{x}{n}$ be the sample proportion of individuals possessing the attribute. Then $E(p) = \frac{1}{n}E(X) = \frac{nP}{n} = P$ and $V(p) = \frac{1}{n^2}\ V(X) = \frac{1}{n^2}\ nP(1-P) = \frac{P(1-P)}{n}$.

For large samples $X \sim N\left[nP,\ nP(1-P)\right]$. Suppose if we want to test the population proportion $P = P_0$, i.e., to test the significant difference between the sample proportion and the population proportion, we set up the statistical hypothesis as

H_0 : There is no significant difference between the sample proportion and the population proportion.

i.e., $H_0 : P = P_0$

Vs

H_1 : There is a significant difference between the sample proportion and the population proportion.

i.e., $H_1 : P \neq P_0$.

The required standard normal test statistic to test the above hypothesis is

$$Z = \frac{p - E(p)}{S.E.(p)} = \frac{P - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \sim N(0,1) \text{ under } H_0.$$

If the calculated value of $|Z|$ is less than the tabulated value of Z at $\alpha\%$ l.o.s., we accept H_0 and conclude that there is no significant difference between the sample proportion and the population proportion. Otherwise reject H_0 and conclude that there is a significant difference between the sample proportion and the population proportion.

12.8 Confidence interval for P:

We have

$$\Pr\left\{|Z| < Z_{\alpha/2}\right\} = 1-\alpha$$

$$\Rightarrow \Pr\left\{\left|\frac{p-P}{\sqrt{\frac{p(1-p)}{n}}}\right| < Z_{\alpha/2}\right\} = 1-\alpha$$

$$\Rightarrow \Pr\left\{|p-P| < Z_{\alpha/2}\left(\sqrt{\frac{p(1-p)}{n}}\right)\right\} = 1-\alpha$$

$$\Rightarrow \Pr\left\{p - Z_{\alpha/2}\left(\sqrt{\frac{p(1-p)}{n}}\right) < P < p + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right\} = 1-\alpha\,,$$

is $100(1-\alpha)\%$ confidence interval for P, and the confidence limits are $p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

For $\alpha = 0.05$, 95% confidence interval for P is given by

$$\Pr\left\{p - 1.96\left(\sqrt{\frac{p(1-p)}{n}}\right) < P < p + 1.96\left(\sqrt{\frac{p(1-p)}{n}}\right)\right\} = 0.95 \text{ and}$$

$p \pm 1.96\left(\sqrt{\frac{p(1-p)}{n}}\right)$ are called confidence limits.

99% confidence interval for P is given by

$$\Pr\left\{p - 2.58\sqrt{\frac{p(1-p)}{n}} < P < p + 2.58\sqrt{\frac{p(1-p)}{n}}\right\} = 0.99$$ and $p \pm 2.58\sqrt{\frac{p(1-p)}{n}}$ are the confidence limits for P.

Problem 1: A die is thrown 9,000 times. A throw of 3 or 4 is observed 3240 times. Test whether the die is unbiased. Also construct 95% confidence interval for P.

Solution: The hypothesis to be tested for the given problem is

H_0 : The die is unbiased

$$\text{i.e.,}\quad H_0 \; : \; P = \frac{1}{3}$$

Vs

H_1 : The die is biased

$$\text{i.e,}\quad H_1 : P \neq \frac{1}{3}$$

Given $n = 9000$; $X = 3240$

$$p = \frac{X}{n} = \frac{3240}{9000} = 0.36$$

Now, the required standard normal test statistic is

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \sim N(0,1)$$

$$Z = \frac{0.36 - 0.33}{\sqrt{\frac{0.33(1-0.33)}{9000}}} = \frac{0.03}{\sqrt{0.0000245}} = 6.06$$

Tabulated value of Z at 5% l.o.s. is 1.96

Since the calculated Z is greater than the tabulated Z at 5% l.o.s., we reject H_0 and conclude that the coin is biased, i.e., $P \neq \frac{1}{3}$.

95% confidence interval for P is given by

$$\Pr\left\{p - 1.96\left(\sqrt{\frac{p(1-p)}{n}}\right) < P < p + 1.96\left(\sqrt{\frac{p(1-p)}{n}}\right)\right\} = 0.95$$

$$\Pr\{0.36 - 1.96(0.004949) < P < 0.36 + 1.96(0.004949)\} = 0.95$$

$$\Pr\{0.36 - 0.0097 < P < 0.36 + 0.0097\} = 0.95$$

$$\Pr\{0.3503 < P < 0.3697\} = 0.95$$

is 95% confidence interval and 95% confidence limits are $(0.3503,\ 0.3697)$.

Problem 2: 20 people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survial rate if attacked by this disease is 85% in favour of this hypothesis that is more at 5% level of significance.

solution: Let P represent the survival rate. The hypothesis to be tested is

H_0 : $P = 0.85$

Vs

H_1 : $P > 0.85$

Also, given $n = 20$; $x = 18$

$$p = \frac{x}{n} = \frac{18}{20} = 0.9$$

The required standard normal test statistic to test the hypothesis is

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{0.9 - 0.85}{\sqrt{\frac{0.85(1-0.85)}{20}}} = 0.626$$

The calculated value of Z at 5% l.o.s. for one tail test is 1.645

Since the calculated value of Z is less than the tabulated value of Z at 5% l.o.s., we accept H_0 and conclude that the survival rate if attacked by the disease is 85%, i.e., $P = 0.85$.

Problem 3: A manufacturer claims that more than 98% of steel pipes supplied to a factory confirms to the specifications. An examination of a sample of 500 pieces of pipes revealed that 30 are defective. Test this claim at 5% l.o.s.

Solution: Let P be the proportion of quality specification or non - defective.

The hypothesis to be tested is

H_0 : We do not support the manufacturer's claim.

i.e., $H_0 : P = 0.98$

Vs

H_1 : We support the manufacturer's claim.

i.e., $H_1 : P > 0.98$

Also, given $n = 500$ and $x = 500 - 30 = 470$

$$p = \frac{x}{n} = \frac{470}{500} = 0.94$$

The required normal test statistic to test the hypothesis is

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0\,(1-P_0)}{n}}} \sim N(0,1)$$

$$Z = \frac{0.94 - 0.98}{\sqrt{\frac{0.98(1-0.98)}{500}}} = -6.35$$

Tabulated value of Z at 5% l.o.s. for one tailtest is 1.645. Since $Z < Z_\alpha$, we accept H_0 and do not support the claim of the manufacturer.

Problem 4: In a sample of 1000 people in Assam 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in the state. Test this at 1% level of significance.

Solution: If both rice and wheat are equally popular in the state, then the probability proportion of rice eaters is $p = \frac{1}{2}$.

The hypothesis to be tested is

H_0 : Both rice and wheat are equally popular in the state.

i.e., $H_0 : P = \frac{1}{2}$

Vs

H_1 : Rice and wheat are not equally popular in the state.

i.e., $H_1 : P \neq \frac{1}{2}$

Given, $n = 1000$ and $x = 540$

$$p = \frac{x}{n} = \frac{540}{100} = 0.54$$

The required standard normal test statistic is

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \sim N(0,1) \text{ under } H_0$$

$$Z = \frac{0.54 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{1000}}} = 2.536$$

Tabulated value of Z at 1% l.o.s. is 2.58

Since, the calculated value of Z is less than the tabulated value of Z at 1% l.o.s., we accept H_0 and conclude that both rice and wheat eaters are equally popular in the state.

12.9 Test of difference of the proportions:

Let X_1 be the number of individuals possessing an attribute in a sample of n_1 individuals drawn from a population with proportion P_1 and let $p_1 = X_1/n_1$.

Let X_2 be the number of individuals possessing the attribute in a sample of n_2 individuals drawn from a population with proportion P_2 and let $p_2 = X_2/n_2$.

Then $p_1 \sim N\left(P_1, P_1(1-P_1)/n_1\right)$

$p_2 \sim N\left(P_2, P_2(1-P_2)/n_2\right)$ for large samples.

$$p_1 - p_2 \sim N\left[P_1 - P_2, \quad \frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}\right] \text{ for large samples.}$$

Suppose, we want to test whether the two population proportions are equal, we set up the statistical hypothesis as

H_0 : There is no significant difference between the two sample proportions.

i.e, $H_0 : P_1 = P_2 = P$

Vs

H_1 : There is a significant difference between the two sample proportions.

i.e., $H_1 : P_1 \neq P_2$

The required standard normal test statistic to test the hypothesis is

$$Z = \frac{p_1 - p_2 - E(p_1 - p_2)}{S.E.(p_1 - p_2)} \sim N(0,1) \text{ under } H_0$$

$$\Rightarrow Z = \frac{p_1 - p_2 - (P_1 - P_2)}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}} \sim N(0,1) \text{ under } H_0$$

$$\Rightarrow Z = \frac{p_1 - p_2}{\sqrt{\frac{P(1-P)}{n_1} + \frac{P(1-P)}{n_2}}} \sim N(0,1) \text{ under } H_0$$

$$Z = \frac{p_1 - p_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Since P is not known in general we will replace it by its pooled estimator

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}.$$

If the calculated value of $|Z|$ is less than the table value of Z at $\alpha\%$ l.o.s., we accept H_0 and conclude that there is no significant difference between the two sample proportions. Otherwise reject H_0.

Problem 1: A company has a head office at Calcutta and a branch at Mumbai. The personal director wanted to know if the workers at the two places would like the introduction of a new plan of work and a survey was conducted for this purpose. Out of a sample of 500 workers at Calcutta 62% favoured the new plan. At mumbai out of sample of 400 workers, 41% were against the new plan. Is there any significant difference between the two groups in their attitude towards the new plan at 5% level.

Solution: Here the hypothesis to be tested is

H_0 : There is no significant difference between the two groups attitude towards the new plan of work.

i.e., $H_0 \ : P_1 = P_2$

Vs

H_1 : There is a significant difference between the two groups attitude towards the new plan of work.

i.e, $H_1 \ : P_1 \neq P_2$

Given $n_1 = 500$; $n_2 = 400$

$p_1 = 62\%$; $p_2 = 100 - 41 = 59\%$

$p_1 = 0.62$; $p_2 = 0.59$

The required standard normal test statistic to test the above hypothesis is

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\ \left(1-\hat{P}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1) \quad \text{under } H_0$$

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{500 \times 0.62 + 400 \times 0.59}{500 + 400} = 0.6$$

$$1 - \hat{P} = 1 - 0.6 = 0.4$$

$$Z = \frac{0.62 - 0.59}{\sqrt{0.6 \times 0.4 \times \left(\frac{1}{500}+\frac{1}{400}\right)}} = 0.917$$

Tabulated value of Z at 5% l.o.s. for two - tailed test is 1.96

Since the caluculated Z is less than the tabulated Z at 5% l.o.s., we accept H_0 and conclude that there is no significant difference between the two groups attitude towards the new plan of work.

Problem 2: A random survey of 400 men and 600 women were asked whether they would like to have fly over near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that the proportions of men and women are same in favour of the proposal at 5% l.o.s.

Solution: Let P_1 be the proportion of men in favour of the proposal and

P_2 be the proportion of women in favour of the proposal.

Here the hypothesis to be tested is

H_0 : The proportions of men and women in favour of the proposal are same

i.e., $H_0 : P_1 = P_2$

Vs

H_1 : The proportions of men and women in favour of the proposal are not same

i.e., $H_1 : P_1 \neq P_2$

Given $p_1 = \frac{X_1}{n_1} = \frac{200}{400} = 0.5$

$$p_2 = \frac{X_2}{n_2} = \frac{325}{600} = 0.54$$

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{525}{1000} = 0.52$$

The required standard normal test statistic to test H_0 is

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\left(1-\hat{P}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \text{ under } H_0$$

$$Z = \frac{0.5 - 0.54}{\sqrt{(0.52)(0.48)\left(\frac{1}{400} + \frac{1}{600}\right)}} = -1.269 \text{ and } |Z| = 1.269$$

Tabulated value of Z at 5% l.o.s. for two tailed test is 1.96

Since the calculated $|Z|$ is less than the tabulated Z at 5% level of significance, we accept H_0 and conclude that the proportions of men and women in favour of the proposal are same.

12.10 Exercises:

1. In a large city A, 20 percent of a random sample of 900 school children had defective eye - sight. In another large city B, 15 percent of random sample of 1600 children had the same defect. Is this difference between the two proportions significant.

2. Before an increase in excise duty on tea, 800 persons out of a sample of 1,000 persons were found to be tea drinkers. After an increase in duty, 800 people were tea drinkers in a sample of 1,200 people. Using standard error of proportion, state whether there is a significant decrease in the consumption of tea after the increase in excise duty?

3. A cigarette manufacturing firm claims that its brand A of the cigarettes outsells its brand B by 8%. It is found that 42 out of a sample of 200 smokers prefer brand A and 18 out of another random sample of 100 smokers prefer brand B. Test whether the 8% difference is a valid claim at 5% l.o.s.

4. On the basis of their total scores 200 candidates of a civil service examination are divided into two groups, the upper 30 percent and the remaining 70 percent. Consider the first question of this examination. Among the first group, 40 had the correct answer, where as among the second group, 80 had the correct answer. On the basis of these results, can one conclude that the first question is no good at discriminating ability of the type being examined here.

5. In a year there are 956 births in a town A, of which 52.5% were males, while in towns A and B combined, this proportion in a total of 1406 births was 0.496. Is there any significant difference in the proportion of male births in the two towns.

6. The mean yield of two sets of plots and their variability are as given below. Examine (i) whether the difference in the mean yields of the two sets of plots is significant and (ii) whether the difference in the variability in yields is significant.

	Set of 40 plots	**Set of 60 plots**
Mean yield per plot	1258 1b	1243 1b
S.D. per plot	34	28

7. In a survey of incomes of two classes of workers, two random samples gave the following details. Examine whether the difference between the standard deviations are significant

Sample	**Size**	**S.D.(in Rs)**
I	100	24
II	100	28

8. A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the S.E. of the population of bad ones in a sample of this size is 0.015 and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5.

9. A coin is tossed 10,000 times and it turns up head 5,195 times. Discuss whether the coin may be regarded as unbiased one.

10. Experience has shown that 20% of a manufactured product is of top quality. In one days production of 400 articles only 50 are of top quality. Show that either the production of the day taken was not a representative sample or the hypothesis of 20% was wrong.

11. A mobile court checking in certain buses it was found that out of 1000 people checked on a certain day at Red Fort 10 persons were found to be ticketless travellers. If daily one lakh passengers travel by the busses, find out the estimated limits to the ticketless travellers.

12.11 Summary:

This lesson covers in detail the second part of large sample tests for variables - Test for single S.D. and test for difference of two S.D.'s. The concept of Fisher's Z transformation and its applications are also mentioned. An attempt of large sample tests for attributes - Test for single proportion and test for difference of two proportions is made with the concept of interval estimation and confidence limits. A number of problems are solved to explain the above concepts and a good number of exercises are given to students to solve on their own.

12.12 Technical Terms:

Test for single standard deviation

Confidence limits

Test for of significance of difference of standard deviations

Sampling of attributes

Test for single proportion

Test for difference of two proportions.

Lesson Writer

C.V. RAO

Lesson 13

SMALL SAMPLE TESTS 1 : TESTS OF SIGNIFICANCE BASED ON t

Objective:

After studying the lesson the students will have clear comprehension of the theory and practical utility of tests of significance based on t - statistic.

Structure of the Lesson:

13.1 Introduction:

If the size of the sample is less than 30, we treat it as small sample. The tests involved for testing any parameter based on the small samples are called small sample tests.

The parameters of normal populations using small samples are tested by test criterian of exact sampling distribution statistics. i.e., t, F and χ^2. This lesson discusses in detail how t - distribution is useful in the following tests of significance.

(i) To test if the sample mean $\left(\bar{x}\right)$ differs significantly from the hypothetical value μ of the population mean.

(ii) To test the significance of the difference between two sample means.

(iii) To test the significance of an observed sample correlation coefficient and sample regression coefficient.

(iv) To test the significance of observed partial correlation coefficient.

Let us discuss each of these in detail.

13.2 t - Test for Single Mean:

Let $x_1, x_2, ..., x_n$ be a random sample of observations drawn from a normal population with mean μ and unknown variance. If one is interested to test the significant difference between the sample mean and the population mean or to test the population mean $\mu = \mu_0$ hypothetically, then we set up the statistical hypothesis as

H_0 : There is no significant difference between sample mean and population mean.

i.e., $H_0 : \mu = \mu_0$

Vs

H_1 : There is a significant difference between sample mean and population mean.

i.e., $H_1 : \mu \neq \mu_0$

The required test statistic to test the above hypothesis is

$$t = \frac{\overline{x} - \mu}{S/\sqrt{n}} = \frac{\overline{x} - \mu}{s/\sqrt{n-1}} \sim t_{n-1} \text{ d.f. at } \alpha \text{ \% l.o.s.}$$

Where $s = \sqrt{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2}$ is a sample S.D.

and $S = \sqrt{\frac{1}{n-1}\Sigma\left(X_i - \overline{X}\right)^2}$ is an unbiased sample S.D.

If the calculated value of $|t|$ is less than the critical value of t with $(n-1)$ d.f. at α % l.o.s., we accept H_0 and conclude that there is no significant difference between sample mean and population mean. Otherwise, we reject H_0 and conclude that there is a significant difference between sample mean and population mean.

Problem 1: A random sample of 10 boys had the following I.Q.'s: 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Does this data support the assumption of population mean I.Q. is 100. Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.

Solution: The hypothesis to be tested is

H_0 : The sample data support the assumption of population mean I.Q. is 100.

i.e., $\mu = 100$

Vs

H_1 : The sample data does not support the assumption of population mean I.Q. is 100.

i.e., $\mu \neq 100$.

Here, the sample mean is

$$\bar{x} = \frac{70+120+110+101+88+83+95+98+107+100}{10}$$

$$\Rightarrow \bar{x} = 97.2$$

and $s = \sqrt{\frac{1}{n}\Sigma\left(X_i - \bar{X}\right)^2}$

X_i :	70	120	110	101	88	83	95	98	107	100
$\left(X_i - \bar{X}\right)^2$:	739.84	519.84	163.84	14.44	84.61	201.64	4.84	0.64	96.04	7.84

$$\Sigma\left(X_i - \bar{X}\right)^2 = 1833.60$$

$$s = \sqrt{\frac{1}{n}\Sigma\left(X_i - \bar{X}\right)^2} = \sqrt{\frac{1}{10}\times 1833.60} = 13.54$$

Now, the required test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} = \frac{97.2 - 100}{13.54/\sqrt{9}} = -0.622$$

$$|t| = 0.62$$

Tabulated value of t for 9 d.f. at 5% l.o.s. for two tailed test is 2.262.

Since the calculated value of t is less than the tabulated value of t, we accept H_0 and conclude that the sample data is consistent with the assumption of population mean $\text{I. Q.} = 100$.

Confidence Limits: The 95% confidence limits within which the population mean I.Q. will lie are given by

$$\bar{x} \pm t_{0.05}\ s/\sqrt{n-1} = 97.2 \pm (2.262)\cdot(4.513) = 97.2 \pm 10.21.$$

Hence the required 95% confidence interval is $[86.99,\ 107.41]$.

Problem 2: The height of 10 soldiers selected at random had a mean height of 158 cm and variance of 39.0625 cms. Assuming a significance level of 5%, test the hypothesis that soldiers of the population are on average less than 162.5 cms tall.

Solution: From the problem, the hypothesis to be tested is

H_0 : The average height of the soldiers is 162.5 cms.

i.e., $H_0 : \mu = 162.5$

Vs

H_1 : The average height of the soldiers is less than 162.50.

i.e, $H_1 : \mu < 162.5$.

Given, $n = 10;\ \bar{x} = 158\ ;\ S^2 = 39.0625$ (sample variance),

the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} = \frac{158 - 162.5}{\sqrt{(39.0625)/9}} = -2.16$$

Since this is a left - tail test, we reject H_0 if $t < -t_\alpha$.

Critical value of t at 5% l.o.s. for 9 d.f. is $-t_\alpha = -1.83$ critical for one - tail test.

Since, computed value of t is less than the critical value of t, H_0 is rejected.

Hence, we conclude that the average height of the soldiers is less than 162.5 cm.

Problem 3: A machinist is making engine parts with axle diameters of 0.700 inches. A random sample of 10 parts shows a mean diameter of 0.742 inches with a standard deviation of 0.040 inches. Compute the statistic you would use to test whether the work is meeting the specifications. Also state how you would proceed further.

Solution: Here, the hypothesis to be tested is

H_0 : The product is conforming to the specified axle diameter.

i.e., $H_0 : \mu = 0.700$

Vs

H_1 : The product is not conforming to the specified axle diameter.

i.e., $H_1 : \mu \neq 0.700$

Also, we are given

$\bar{x} = 0.742$ inches, $s = 0.040$ and $n = 10$

The required test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} = \frac{0.742 - 0.700}{0.040/\sqrt{9}} = 3.15\,.$$

The tabulated value of t at 5% l.o.s. for 9 d.f. is 2.262

Since, the calculated value of t is greater than the tabulated value of t we reject the hypothesis and conclude that the product is not conforming to the specifications of the axle diameter.

13.3 t - test for difference of two means:

Let $x_1, x_2, \cdots\cdots\cdots, x_i, \cdots\cdots, x_{n_1}$ be a random sample taken from the normal population with mean μ_1 and unknown variance. Let $\bar{x}$ be the sample mean and s_1^2 be the sample variance. Let S_1^2 be the unbiased estimated of population varience, where $\bar{x} = \frac{\sum X_i}{n_1}$ and

$$s_1^2 = \frac{1}{n_1}\Sigma\left(X_i - \bar{X}\right)^2, \qquad S_1^2 = \frac{1}{n_1 - 1}\Sigma\left(X_i - \bar{X}\right)^2.$$

Let $y_1, y_2, \cdots\cdots\cdots, y_j, \cdots\cdots\cdots, y_{n_2}$ be another random sample taken from the normal population with mean μ_2 and unknown variance. Let $\bar{y}$ be the sample mean and s_2^2 be the sample variance. Let S_2^2 be the unbiased estimate of population variance. Let $\bar{Y} = \frac{\sum Y_j}{n_2}$, $s_2^2 = \frac{1}{n_2}\Sigma\left(y_j - \bar{y}\right)^2$, $S_2^2 = \frac{1}{n_2 - 1}\Sigma\left(y_j - \bar{y}\right)^2$. Now if one is interested to test the significance difference between the two sample means $\bar{x}$ and $\bar{y}$ or the equality of the two population means μ_1 and μ_2, we need to set up the null and alternative hypothesis as

H_0 : There is no significance difference between two sample means.

i.e., $H_0 : \mu_1 = \mu_2$

Vs

H_1 : There is a significance difference between the two sample means.

i.e., $H_1 : \mu_1 \neq \mu_2$

The required test statistic to test the above hypothesis is

$$t = \frac{\left(\bar{x} - \bar{y}\right) - \left(\mu_1 - \mu_2\right)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{\left(n_1+n_2-2\right)}; \infty$$

where $S^2 = \frac{\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_2^2}{n_1 + n_2 - 2}$

or $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

Since $\mu_1 = \mu_2$ under H_0, t reduces to

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{\left(n_1+n_2-2\right)\text{d.f.}} \quad \text{at } \alpha\% \quad \text{l.o.s.}$$

If the calculated value of $|t|$ less is than the tabulated value of t with $\left(n_1 + n_2 - 2\right)$ d.f. and at $\alpha\%$ l.o.s., we accept H_0, and conclude that the two population means are equal. Otherwise, we reject H_0 and conclude that there is a signigicance difference between the two sample means.

Problem 1: The number of sales, average size of the sales and standard deviation of two sales persons are as follows:

	A	**B**
Number of Sales	10	17
Average size (in Rs)	6,200	5,600
Standard deviation (in Rs)	690	600

Examine whether the figures of average sales size differ significantly at 5% l.o.s.

Solution: The hypothesis to be tested is

H_0 : The average sales figures of the two persons A and B does not differ significantly.

i.e., $H_0 : \mu_1 = \mu_2$

Vs

H_1 : There is a significance difference in the sales figures of the two persons.

i.e., $H_1 = \mu_1 \neq \mu_2$.

Also, from the problem we have

$n_1 = 10$; $n_2 = 17$

$\bar{x} = 6,200$; $\bar{y} = 5,600$

$s_1 = 690$; $s_2 = 600$

The required test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Now, $$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{10(690)^2 + 17(600)^2}{10 + 17 - 2}$$

$$S = \sqrt{435240} = 659.73$$

$$t = \frac{6200 - 5600}{659.73\sqrt{\frac{1}{10} + \frac{1}{17}}} = \frac{600}{262.92} = 2.28206$$

Tabulated value of t at 5% l.o.s. for 25 d.f. is 2.06

Since, the calculated value of t is greater than the tabulated value, we reject H_0 and conclude that there is a significance difference in the average sales of two sales persons A and B.

Problem 2: Measurments of fat contents of two brands of ice- creams A and B yielded the following data

Brand A	**:**	13.5,	14,	13.6,	12.9,	30
Brand B	**:**	12.9,	13,	12.4,	13.5,	12.7

Test whether average fat contents of the two brands of ice - creams differ significantly at 1% l.o.s.

Solution: The hypothesis to be tested is

H_0 : There is no signigicant difference in the average fat content of two brands of ice - creams.

i.e, $H_0 \; : \mu_1 = \mu_2$

Vs

H_1 : There is a significance difference in the average fat content of two brands of ice - creams.

The required test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where, $\bar{x} = \frac{\sum x_i}{n_1}$

$$\bar{y} = \frac{\sum y_j}{n_2}$$

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \quad ; \quad s_1^2 = \frac{1}{n_1} \sum\left(x_i - \bar{x}\right)^2$$

$$s_2^2 = \frac{1}{n_2} \sum\left(y_j - \bar{y}\right)^2$$

x	y	$\left(x-\bar{x}\right)^2$	$\left(y-\bar{y}\right)^2$
13.5	12.9	10.89	0
14.0	13.0	7.84	0.01
13.6	12.4	10.24	0.25
12.9	13.5	15.21	0..36
30.0	12.7	174.24	0.04
84.0	64.5	218.42	0.66

$$\bar{x} = \frac{84.0}{5} = 16.8$$

$$\bar{y} = \frac{64.5}{5} = 12.9$$

$$s_1^2 = \frac{1}{5} \times 218.42 \quad ; \quad s_2^2 = \frac{1}{5} \times 0.66$$

$$s_1^2 = 43.684 \quad ; \quad s_2^2 = 0.132$$

$$S^2 = \frac{5(43.684)+5(0.132)}{5+5-2} = \frac{218.42+0.66}{8} = 27.385$$

$$S = 5.23306$$

Now,

$$t = \frac{16.8-12.9}{5.23306\sqrt{\frac{1}{5}+\frac{1}{5}}} = \frac{3.9}{3.30968} = 1.17836$$

Tabulated value of t with 8 d.f. at 1% l.o.s. is 3.36.

Since, the calculated value of t is less than the tabulated value of t with 8 d.f. at 1% l.o.s., we accept H_0, and conclude that there is no significant difference in the average fat content of two brands of ice - creams.

13.4 Paired t - test for difference of two means:

In t - test for difference of two population means, we have taken two independent random samples from two normal populations and interested to test the equality of two population means.

Now, this is an another situation, if the two samples are not independent and sizes of these two samples are equal and we want to test the significant difference between the two sample means, we have to apply paired - t - test.

Let $x_1, x_2, \cdots\cdots\cdots\cdots, x_n$ and $y_1, y_2, \cdots\cdots\cdots\cdots, y_n$ be two dependent samples of size 'n' and if one is interested to test the significance difference between these two sample means, we set up the statistical hypothesis as

H_0 : There is no significant difference between the two sample means.

Vs

H_1 : There is a significant difference between the two sample means.

Here, we consider $d_i = (x_i - y_i) \; \forall \; i = 1, \ldots\ldots\ldots, n$. The required t - test to test the above hypothesis is

$$t = \frac{\bar{d}}{s/\sqrt{n-1}} \sim t_{(n-1)} \text{ d.f}$$ at α % l.o.s.

Where , $\bar{d} = \frac{\sum d_i}{n}$

$$s = \sqrt{\frac{1}{n}\sum\left(d_i - \bar{d}\right)^2}.$$

If the calculated value of $|t|$ is less than the tabulated value of t with $(n-1)$ d.f. and at α% l.o.s., we accept H_0 and conclude that there is no significant difference between the two sample means, otherwise we reject H_0 and conclude that there is a significant difference between the two sample means.

Problem 1: In a certain experiment to compare two types of foods A and B applied on the same set of 8 cows, the following results of increase in wights were observed.

Food A :	49	53	51	52	47	50	52	53
Food B :	52	55	52	53	50	54	54	53

Can you conclude that Food B is superior to Food A.

Solution: Since, the samples are not independent we have to apply paired t - test to test the above problem.

From the given problem the statistical hypothesis to be tested is

H_0 : There is no significant difference in the Foods of A and B regarding to gain in weights.

Vs

H_1 : The Food B is superior to Food A w.r.t. gain in weights (one - tail test)

We construct the test statistic as

$$t = \frac{\bar{d}}{s/\sqrt{n-1}}$$

Food A	**Food B**		
X	**Y**	$d = (X-Y)$	$(d-\bar{d})^2$
49	52	-3	1
53	55	-2	0
51	52	-1	1
52	53	-1	1
47	50	-3	1
50	54	-4	4
52	54	-2	0
53	53	0	4
		$\sum d = -16$	12

$$\bar{d} = \frac{\sum d_i}{n} = \frac{-16}{8} = -2$$

$$s^2 = \frac{1}{n}\sum\left(d_i - \bar{d}\right)^2 = \frac{12}{8} = 1.5 \quad ; \quad s = \sqrt{1.5} = 1.2247$$

Now,

$$t = \frac{-2}{(1.2247)/\sqrt{8-1}} = \frac{(-2)\left(\sqrt{7}\right)}{1.2247}$$

$$t = \frac{-5.2915}{1.2247} = -4.32065$$

$$|t| = 4.32$$

The table value of t for 7 d.f. at 5 % l.o.s. for one - tail test is 1.9

Since, the calculated value of t is greater than the tabulated value of t for 7 d.f. at 5 % l.o.s., we reject H_0 and conclude that Food B is superior to Food A.

Problem 2: A certain stimulus administered to each of 12 patients resulted in the following increase of blood pressure:

5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4 and 6.

Can it be concluded that the stimulus will, in general, be accompained by an increase in blood pressure?

Solution: From the problem the hypotheses to be tested is

H_0 : There is no charge in the blood pressure by stimulus administered

Vs

H_1: There is an increase in the blood pressure by the stimulus adminisered.

Given thevalues of d_i

$$\therefore \bar{d} = \frac{5+2+8-1+3+0-2+1+5+0+4+6}{12} = \frac{31}{12} = 2.58$$

Also, we need to calculate s

d :	5	2	8	-1	3	0	-2
$(d-\bar{d})^2$:	5.8564	0.3364	29.3764	2.4964	0.1764	6.6564	0.3364
d :	1	5	0	4	6		
$(d-\bar{d})^2$:	2.4964	5.8564	6.6564	2.0164	11.6964		

$$\Sigma(d-\bar{d})^2 = 73.9568$$

$$s = \sqrt{\frac{1}{n}\Sigma(d-\bar{d})^2} = \sqrt{\frac{1}{12}(73.9568)} = 2.4825$$

Now, the test statistic is

$$t = \frac{2.58}{2.4825/\sqrt{11}} = \frac{8.5568}{2.4825} = 3.4$$

Tabulated value of t for 11 d.f. at 5% l.o.s. for one - tail test is 1.80.

Since, the calculated value of t is greater than the tabulated value of t for 11 d.f. at 5% l.o.s., we reject H_0 and conclude that there is an increase in bold pressure accompanied by the stimulus administered.

Problem 3: Two laboratories carry out independent estimates of a particular chemicals in a medicine produced by a certain firm. A sample is taken from each batch, halved and the seperate halves sent to the two laboratories. The following data is obtained:

No. of samples : 10

Mean value of difference of estimates : 0.6

Sum of squares of the difference (from the mean) : 20

Is the difference in the two laboratories estimates is significant. Comment on your findings at 5% l.o.s.

Solution: The hypothesis to be tested is

H_0 : The difference is insignificant

Vs

H_1 : The difference is significant

We are given n = 10,

$\bar{d} = 0.6$ and $\sum\left(d-\bar{d}\right)^2 = 20$

$$s = \sqrt{\frac{1}{n}\Sigma\left(d-\bar{d}\right)^2} = \sqrt{\frac{1}{10}\times 20} = 1.4142$$

Now, the required test statistic is

$$t = \frac{\bar{d}}{s/\sqrt{n-1}} = \frac{(0.6)\sqrt{9}}{1.4142} = \frac{1.8}{1.4142} = 1.272$$

Tabulated value of t for 9 d.f. at 5% l.o.s. is 2.262

Since, the calculated value of t is less than the tabulated value of t for 9 d.f. at 5% l.o.s., we accept H_0 and conclude that the difference in the estimates of chamicals made by the two tabs is insignificant.

13.5 t - test for testing the significance of sample correction coefficient:

Let $(x_i, y_i) \forall i = 1, 2, ..., n$ be a bi - variate random sample drawn from a bi - variate normal population with the correlation coefficient $'\rho'$.

Let $r = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$ be the sample correlation coefficient. If we are interested to test whether the population correlation coefficient $\rho = 0$, i.e., the two variables are uncorrelated, we need to set up the following null and alternative hypothesis.

H_0 : The variables are uncorrelated

i.e., $\rho = 0$

Vs

H_1 : The variables are correlated

i.e., $\rho \neq 0$

Fisher proved that under H_0, the statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)} \text{ d.f.}$$

If the calculated value of $|t|$ is less than the tabulated value of t for $(n-2)$ d.f. at α % l.o.s, we accept H_0 and conclude that the variables are uncorrelated, otherwise reject H_0 and say that the two variables are correlated.

Problem 1: (a) A random sample of 27 pairs of observations from a normal population gave a correlation coefficient of 0.6. Is this significant of correlation in the population?

(b) Find the least value of r in a sample of 18 pairs of observations from a bivariate normal population, significant at 5% l.o.s.

Solution: (a) The hypothesis to be tested is

H_0 : Sample correlation coefficient is not significant in the population.

i.e., $H_0 : \rho = 0$

Vs

H_1 : Sample correlation coefficient is significant in the population.

i.e., $H_1 : \rho \neq 0$

Given $n = 27$; $r = 0.6$

Now, the required test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.6\sqrt{27-2}}{\sqrt{1-(0.6)^2}} = \frac{3}{\sqrt{0.64}} = 3.75$$

Tabulated value of t for 25 d.f. at 5% l.o.s. is 2.06

Since, the calculated value of t greater than the tabulated value of t at 5% l.o.s. for 25 d.f., we reject H_0 and conclude that sample correlation coefficient is significant in the population. That is the variables are correlated in the population.

(b) Given n = 18; also tabulated value of t at 5% l.o.s. for $(n-2) = 16$ d.f. is 2.12. It is given that r is significant at 5% level of significance and the test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{(1-r)^2}} \sim t_{(n-2)}$$

$$\Rightarrow |t| > t_{0.05}, (n-2)$$

$$\Rightarrow \left|\frac{r\sqrt{n-2}}{\sqrt{(1-r)^2}}\right| > 2.12$$

$$\Rightarrow \left|\frac{r\sqrt{16}}{\sqrt{1-r^2}}\right| > 2.12$$

$$\Rightarrow 16r^2 > (2.12)^2 \left(1-r^2\right)$$

$$\Rightarrow 16r^2 + 4.4944r^2 > 4.4944$$

$$\Rightarrow 20.4944r^2 > 4.4944$$

$$\Rightarrow r^2 > \frac{4.4944}{20.4944}$$

$$\Rightarrow \; r^2 > 0.2192$$

$$\Rightarrow \; |r| > 0.4682.$$

Problem 2: A correlation coefficient of 0.2 is derived from a random sample of 625 pairs of observations. Is this value of r significant at 5% level of significance.

Solution: Hypothesis to be tested is

H_0 : Sample correlation coefficient $r = 0.2$ is not significant.

i.e., $H_0 \; : \rho = 0$

Vs

H_1 : Sample correlation coefficient is significant.

i.e. $H_1 \; : \rho \neq 0$

The reequired test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} \quad \text{at } \alpha\% \text{ l.o.s.}$$

Given $n = 625 \; ; \; r = 0.2$

$$t = \frac{0.2\sqrt{625-2}}{\sqrt{1-(0.2)^2}} = \frac{0.2(24.9599)}{0.9797} = \frac{4.9919}{0.9797} = 5.09$$

Now, d.f. here are $(n-2) = 623$. Since sample size is large, significant values of t are same as in the case of normal distribution.

We know that the critical value of Z at 5% l.o.s. is 1.96 in a two tailed test. Since the calculated value of t is greater than the critical value, we reject H_0 and conclude that the sample corrrelation coefficient is significant in the population.

13.6 t - test for testing the significance of an observed Regression Coefficient:

Here the problem is to test if a random sample (x_i, y_i), $(i = 1, 2, ..., n)$ has been drawn from a bivariate normal population in which regression coefficient y on x is β.

The line of regression of y on x is

$$y - \bar{y} = b\left(x - \bar{x}\right), \quad b = \frac{\mu_{11}}{\sigma_x^2}$$

From this line the estimate of y for a given value x_i of x is given by

$$\hat{y}_i = \bar{y} + b\left(x_i - \bar{x}\right).$$

Under the null hypothesis H_0 that the population regression coefficient is β, Prof R.A. Fisher proved that the statistic,

$$t = \left(b - \beta\right)\left[\frac{\left(n-2\right)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}\right]^{\frac{1}{2}}$$

follows Student's t - distribution with $\left(n-2\right)$ d.f.

13.7 t - test for testing the significance of an observed partial correlation coefficient:

Let $r_{12.34............(k+2)}$ be the partial correlation coefficient of order k, observed in a sample of size n from a multivariate normal population, Prof. Fisher proved that under the null hypothesis $H_0 : \rho_{12.34............(k+2)} = 0$, i.e., the population partial correlation coefficient is zero, the statistic

$$t = \frac{r_{12}}{\sqrt{1 - r_{12}^2}}\sqrt{n - k - 2}$$

follows Student's t - distirbution with $\left(n - k - 2\right)$ d.f.

13.8 Exercises:

1. A random sample of 9 experimental animals under a certain diet gave the following increase in weight : $\sum x_i = 45$ lbs, $\sum x_i^2 = 279$ lbs, where x_i denotes that increase in weight of the ith animal. Assuming that the increase in weight is normally distributed as $N\left(\mu, \sigma^2\right)$ variate, test $H_0 : \mu = 1$ against $H_1 : \mu \neq 1$ at 5% l.o.s.

2. A random sample of eight cigarettes of a certain brand has an average nicotine content of 18.6 milligrams and a standard deviation of 2.4 milligrams. In this line

with the manufactures claim that average nicotine content does not exceed 17.5 milligrams. Use 0.01 level of significance and assume the distribution of nicotine contents to be normal.

3. The average length of time for students to register for summer classes at a certain college has been 50 minutes with s.d. of 10 min. A new registration procedure using modern computing machines is being tried. If a random sample of 12 students had an average registration time of 42 minutes with s.d. of 11.9 minutes under the new system, test the hypothesis that the population mean has not changed, using 0.05 as level of significance.

4. The average breaking strength of steel rod is 18.5 thousand pounds. To test this a sample of 14 rods was tested. The mean and standard deviations obtained were 17.85 and 1.9555 thousand pounds respectively. Is the result of the experiment significant? Also obtain 95% fiducial limits from the sample for the average breaking strength of steel rods.

5. A random sample of 8 envelops is taken from letter box of a post office and their weights in grams are found to be 12.1, 11.9, 12.4, 12.3, 11.9, 12.1, 12.4, 12.1. Test whether the average weight of envelopes received at that post office is 12.35 gms. Find 99% confidence limits for the mean weight of the envelopes received at the post office.

6. Two horses A and B were tested according to time (in seconds) to run a particular track with the following results:

Horse A:	13	14	10	11	12	16	10	8	
Horse B:	7	10	12	8	10	11	9	10	11

Test whether the two horses have the same running capacity at both 5% and 1% l.o.s.

7. Show how you would use students t - test to decide whether the two sets of observations [17, 27, 18, 25, 27, 29, 27, 23, 17] and [16, 16, 20, 16, 20, 17, 15, 21] indicate samples drawn from the same universe.

8. Ten soldiers visit a riffle range for two consecutive weeks, which are not independent and the scores for the two weeks are 67, 24, 57, 55, 63, 54, 56, 68, 33, 43 and 70, 38, 58, 58, 56, 67, 68, 72, 42, 38 respectively. Examine if there is any significant difference in their performance at your desired level of significance.

9. A random sample of 16 values from a normal population has a mean of 41.5 inches and sum of squared of deviations from mean is equal to 135 inches. Another sample of 20 values from an unknown population has a mean of 43.0 inches and sum of squares of deviations from their mean is equal to 171 inches show that two samples may be regarded as coming from the same normal population.

10. The following table shows the mean number of bacterial colonies per plate obtainable by four slightly different methods from soil samples taken at 4 p.m. and 8 p.m.

respectively:

Method	**A**	**B**	**C**	**D**
4 P.M.	29.75	27.50	30.25	27.80
8 P.M.	39.20	40.60	36.20	42.40

11. An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material I were tested, by exposing each piece to a machine measuring wear. Ten pieces of material II were similary tested. In each case thedepth of wear was observed. The sample of material I gave an average (coded) wear 8.5 units with a standard deviation of 0.4 white the sample of material II gave an average of 8.1 and a standard deviation of 0.5. Test the hypothesis that the two types of material exhibit the same mean abrasive wear at the 0.01 l.o.s. Assume the populations to be approximately normal with equal variances.

12. A random sample of 27 pairs of observations from a normal population gives a correlation coefficient of 0.42. Is it likely that the variables in the population are uncorrelated.

13. Find the least value of r in a sample of 27 pairs from a bivariate normal population significant at 5% level of significance.

13.9 Summary:

Various applications of t - test are discussed in detail. A good number of problems are solved and a number of exercises are given to students to solve on their own.

13.10 Technical Terms:

Test for single mean

Test for difference of means

Paired t - test

Test for observed sample correlation coefficient

Lesson Writer

P. Nagamani

Lesson 14

SMALL SAMPLE TESTS 2 : TESTS OF SIGNIFICANCE BASED ON χ^2 AND F DISTRIBUTIONS

Objective:

After studying this lesson the students will have a clear understanding of theoritical concepts as well as practical utility of tests of significance based on χ^2 and F distributions.

Structure of the Lesson:

14.1 Introduction:

In the previous lesson we have discussed the applications of a small sample test based on t - distribution. Now we will continue to discuss the applications of small sample tests based on χ^2 and F distributions in this lesson.

14.2 χ^2 - test for significance of Normal Population Variance:

Let us consider a normal population with mean μ and variance σ^2. Let $x_1, x_2, ..., x_n$ be a random sample drawn from the normal population, then $\chi^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$ is a χ^2 - variate with n d.f. when the population mean μ is known.

If population mean μ is not known replace μ by $\overline{X}$ and now the χ^2 variate is $\chi^2 = \sum_{i=1}^{n}\left(\frac{X_i - \overline{x}}{\sigma}\right)^2 \sim \chi^2_{n-1}$ d.f. Suppose, we are interested to test the population variance $\sigma^2 = \sigma_0^2$ or significant difference between the population variance and the sample variance, we setup the statistical hypothesis as

H_0 : There is no significant difference between the population variance andthe sample variance

i.e., $H_0 : \sigma^2 = \sigma_0^2$.

Vs

H_1 : There is a significant difference between the population variance and the sample variance.

i.e., $H_1 : \sigma^2 \neq \sigma_0^2$.

Under H_0, the required test statistic to test the above hypothesis when μ is not known is

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{X_i - \overline{x}}{\sigma_0}\right)^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma_0^2} = \frac{ns^2}{\sigma_0^2} \sim \chi^2_{(n-1)} \text{ d.f.}$$

at $\alpha\%$ l.o.s., where $s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}$

If the calculated value of χ^2 is less than the tabulated value of χ^2 for $(n-1)$ d.f. at $\alpha\%$ l.o.s., we accept the hypothesis and conclude that there is no significant difference between the population variance and the sample variance. Otherwise, reject the hypothesis and conclude that there is a significance difference between the population variance and the sample variance.

Note: Particularly for large samples, i.e., $n > 30$, we have Fisher's approximation $\sqrt{2\chi^2} \sim N\left(\sqrt{2n-1}, 1\right)$. And if we are interested to test the above hypothesis the required test statistic is

$$Z = \left[\sqrt{2\chi^2} - \sqrt{2n-1}\right] \sim N(0,1) \text{ under } H_0$$

and the test procedure is the usual normal test.

Problem 1: A random sample of size 20 from a population gives the sample standard deviation of 6. Test the hypothesis that the population standard deviation is 9.

Solution: Here, we set up the null and alternative hypothesis as

$$H_0 : \sigma^2 = (9)^2$$

Vs

$$H_1 : \sigma^2 \neq (9)^2$$

Also given $n = 20$ and sample standard deviation is 6.

i.e., $s^2 = 36$. Now, the required test statistic is

$$\chi^2 = \frac{ns^2}{\sigma_0^2} = \frac{20 \times (6)^2}{(9)^2} = 8.89$$

Tabulated value of χ^2 for $(20-1) = 19$ d.f. at 5% l.o.s. is 30.144

Since, the calculated value of χ^2 is less than the tabulated value of χ^2 for 19 d.f. at 5% l.o.s., we accept H_0 and conclude that the population standard deviation is 9.

Problem 2: Weights in Kg. of 10 students, ae given below:

38, 40, 45, 53, 47, 43, 55, 48, 52, 49

Can we say that the variance of distribution of weights of all students from which the above sample of 10 students was drawn, is equal to 20 Kg?

Solution: Here, we setup the null and alternative hypothesis as

$$H_0 : \sigma^2 = 20$$

Vs

$$H_1 : \sigma^2 \neq 20$$

The required test statistic is

$$\chi^2 = \frac{\sum_{i=1}^{10} \left(X_i - \overline{X}\right)^2}{\sigma_0^2}$$

Table for computation of sample variance

weights (Kgm) X	$\left(X-\overline{X}\right)$	$\left(X-\overline{X}\right)^2$
38	-9	81
40	-7	49
45	-2	4
53	6	36
47	0	0
43	-4	16
55	8	64
48	1	1
52	5	25
49	2	4
$\sum x = 470$		$\sum\left(x-\overline{x}\right)^2 = 280$

$$\overline{x} - \frac{\sum x}{n} = \frac{470}{10} = 47$$

Now,

$$\chi^2 = \frac{\sum\left(X_i-\overline{X}\right)^2}{\sigma_0^2} = \frac{280}{20} = 14$$

Tabulated value of χ^2 for 9 d.f. at 5% l.o.s. is 16.92

Since, the calculated value of χ^2 is less than the tabulated value of χ^2 for 9 d.f. at 5% l.o.s., we accept H_0 and conclude that the variance of distribution of weights of all students in the population is 20 Kg.

Problem 3: Test the hypothesis that $\sigma = 8$ given that $s = 10$ for a random sample of size 51.

Solution: The null and alternative hypothesis in this problem is

$H_0 : \sigma = 8$

Vs

$H_1 : \sigma \neq 8$

Also, we are given $n = 51$ and $s = 10$

The required test statistic is

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{51 \times 100}{64} = 79.69$$

Since, n is large we can use Fisher's normal approximation to chi - square i.e., the test statistic is

$$Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0,1)$$

$$Z = \sqrt{2 \times 79.69} - \sqrt{102-1}$$

$$Z = 2.57$$

We know that tabulated value of Z at 5% l.o.s. for two - tailed test is 1.96

Since, the calculated value of Z is greater than the tabulated value of Z, we reject H_0 and conclude that random sample is drawn from a population whose S.D. is not 8.

14.3 χ^2 - test of Independence of Attributes:

Let us consider two attributes A and B, where attribute A is divided into r classes say $A_1, A_2, ..., A_r$ and attribute B be divided into s classes say $B_1, B_2, ..., B_s$. The occurrance of both the attributes in a population of N individuals is represented in the following table called contingency table of $r \times S$ $(\text{rows} \times \text{coloumns})$.

	B_1	B_2	....	B_j	....	B_s	
A_1	(A_1B_1)	(A_1B_2)	....	(A_1B_j)	....	(A_1B_s)	(A_1)
A_2	(A_2B_1)	(A_2B_2)	....	(A_2B_j)	....	(A_2B_s)	(A_2)
⋮	⋮						
A_i	(A_iB_1)	(A_iB_2)	....	(A_iB_j)	....	(A_iB_s)	(A_i)
⋮	⋮						⋮
A_r	(A_rB_1)	(A_rB_2)	....	(A_rB_j)	....	(A_rB_s)	(A_r)
	(B_1)	(B_2)	....	(B_j)	....	(B_s)	N

where,

$\left(A_iB_j\right)$ is the observed frequency representing the number of individuals possessing both the attributes A_i and B_j $\forall$ $i=1,2,..........,r$; $j=1,2,.........,s$.

$(A_i) = \sum\limits_{j=1}^{s}\left(A_iB_j\right)$ which indicates the number of individuals posessing the attribute A_i $\forall$ $i=1,2,.............,r$.

$\left(B_j\right) = \sum\limits_{i=1}^{r}\left(A_iB_j\right)$ which indicates the number of individuals posessing the attribute B_j $\forall$ $j=1,2,...............,s$.

Also let $\left(A_iB_j\right)_0 = \dfrac{(A_i)\left(B_j\right)}{N}$ is the expected frequency for $\left(A_iB_j\right)$ $\forall$ $i=1,2,......,r$ and $j=1,2,............,s$.

Now, if we are interested to test whether the two attributes A and B are independent or associated, then we set up the statistical hypothesis as

H_0 : The two attributes A and B are independent.

Vs

H_1 : The two atributes A and B are associated.

The required test statistic to test the above hypothesis is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left\{\frac{\left[\left(A_iB_j\right)-\left(A_iB_j\right)_0\right]^2}{\left(A_iB_j\right)_0}\right\} \sim \chi^2_{(r-1)(s-1)} \text{ d.f. } \alpha \text{ \% l.o.s.}$$

If the calculated value of χ^2 is less than the tabulated value of χ^2 for $(r-1)(s-1)$ d.f. at α % l.o.s., we accept H_0 and conclude that the two attributes A and B are independent, otherwise we reject H_0 and conclude that the two attributes A and B are associated.

Note: For two attributes A and B, each divided into two classes A_1, A_2 and B_1, B_2 the 2×2 contingency table is

	B_1	B_2	
A_1	a	b	(A_1)
A_2	c	d	(A_2)
	(B_1)	(B_2)	N

Now, the required test statistic to test the hypothesis is

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \sim \chi^2_{(2-1)(2-1)} \text{ d.f. at } \alpha\% \text{ l.o.s.}$$

where $N = a+b+c+d$

In case of 2×2 contingency table if any one of the frequency is less than five, Yates correction of continuity is to be made. For this the value of 0.5 is to be added for the frequency less than 5 and then adjusted to the rest of the frequencies so that total does not change. Then the χ^2 - test is given by

$$\chi^2 = \frac{N\left[|ad-bc| - N/2\right]^2}{(a+b)(a+c)(b+d)(c+d)}.$$

However, it is recommended to use Yates correction to every 2×2 contingency table, even if no theoritical cell frequency is less than 5.

Problem 1: Suppose that, in a public opinion survey answers to the questions -

(a) Do you drink?

(b) Are you in favour of local opinion on sale of liquor?

were as tabulated below:

	Question (a)		**Total**
Question (b)	**Yes**	**No**	
Yes	56	31	87
No	18	6	24
Total	74	37	111

Can you infer that opinion on local option is dependent on whether or not an individual drinks.

Solution: Let A represent an individual drink or not drink

B represent an individual in favour of local option on sale of liquor.

Now, the hypothesis to be tested is

H_0 : Two attributes under consideration are independent.

Vs

H_0 : Two attributes ae associated.

From the given data $a = 56$; $b = 31$; $c = 18$; $d = 6$

Yates correction of χ^2 for 2×2 contingency table for the above hypothesis is

$$\chi^2 = \frac{N\left[|ad - bc| - N/2\right]^2}{(a+b)(a+c)(b+d)(c+d)}$$

$$\chi^2 = \frac{111\left[|336 - 558| - 55.5\right]^2}{(87)(74)(37)(24)}$$

$$\chi^2 = 1.49$$

Here the degrees of freedom is $(2-1) \times (2-1) = 1$ d.f.

Tabulated value of χ^2 for 1 d.f. at 5% l.o.s. is 3.841.

Since the calculated χ^2 is less than the tabulated χ^2 for 1 d.f. at 5% l.o.s., we accept H_0 and conclude that the two attributes under consideration are independent, i.e., the opinion on local option on sale of liquor is not dependent on whether or not an individual drinks.

Problem 2: In an experiment on the immunization of goats from Anthrax, the following results were obtained. Derive your inference on the efficiency of the vaccine.

	Died of Anthrax	**Survived**	**Total**
Innoculated with vaccine	2	10	12
Not innoculated	6	6	12
Total	8	16	24

Solution: Let A represent innoculation with vaccine

B represent survival from anthrax

Here, hypothesis to be tested is

H_0 : The two attributes are independent

Vs

H_1 : The two attributes are associaged.

Here the hypothesis is tested with Yates correction of χ^2 for 2×2 contingency table. Here one of the frequency is less than 5 hence we add 0.5 to that frequency and modify the frequencies accordingly.

			Total
	2.5 5.5	9.5 6.5	12 12
Total	8	16	24

Yates correction of χ^2 for 2×2 contingency table for the above hypothesis is given by

$$\chi^2 = \frac{N\left[|ad - bc| - N/2\right]^2}{(a+b)(a+c)(b+d)(c+d)}$$

$$= \frac{24\left[|(2.5)(6.5) - (9.5)(5.5)| - 24/2\right]^2}{(2.5+9.5)(2.5+5.5)(9.5+6.5)(5.5+6.5)}$$

$$\chi^2 = 0.75$$

Tabulated value of χ^2 for 1 d.f. at 5% l.o.s. is 3.841

Since, the calculated value of χ^2 is less than the tabulated χ^2 value, we accept H_0 and conclude that the two attributes are independent, i.e., innoculation with vaccince has no effect on survival from Anthrax.

Problem 3: 1072 college students were classified according to their intelligence and economic conditions. Test whether their is association between intelligence and economic conditions.

		Intelligence			
		Excellent	Good	Medium	Dull
Economic Conditions	Good	48	199	181	82
	Not Good	81	185	190	106

Solution: Let A represent economic conditions of the students

B represent intelligence of the students

Hence, the hypothesis to be tested is

H_0 : The two attributes are independent.

Vs

H_1 : The two attributes are assosiated.

χ^2 statistic for testing the above hypothesis is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left\{\frac{\left[(A_iB_j)-(A_iB_j)_0\right]^2}{(A_iB_j)_0}\right\} \sim \chi^2_{(r-1)(s-1)\text{ d.f.}} \text{ at } \alpha\% \text{ l.o.s.}$$

		Intelligence (B)				
		Excellent	Good	Medium	Dull	Total
Economic	Good - A_1	48	199	181	82	510
Conditions (A)	Not Good A_2	81	185	190	106	562
	Total	129	384	371	188	1072

(A_iB_j)	$(A_iB_j)_0$	$\left[(A_iB_j)-(A_iB_j)_0\right]^2$	$\left[(A_iB_j)-(A_iB_j)_0\right]^2 / (A_iB_j)_0$
48	61.37	178.7569	2.9127
199	182.68	266.3424	1.4579
181	176.50	20.2500	0.1147
82	89.44	55.3536	0.6188
81	67.62	179.0244	2.6475
185	201.31	266.0161	1.3214
190	194.49	20.1601	0.1036
106	98.55	55.5025	0.5631
			9.7401

$\therefore \chi^2 = 9.7401 \sim \chi^2_{(2-1)(4-1)}$ at 5% l.o.s.,

Tabulated value of χ^2 for 3 d.f. at 5% l.o.s. is 7.815

Since the calculated value of χ^2 is greater than the tabulated value of χ^2 for 3 d.f. at 5% l.o.s., We reject H_0 and conclude the two attributes are associated i.e., there is an association between intelligence and economic conditions.

14.4 χ^2 - test of goodness of fit:

χ^2 - test of goodness of fit helps us to study the deviation between observed (experimental) and expected (theoritical) frequencies. This was developed by Prof. Karl Pearson in 1900. The following is the procedure of χ^2 test of goodness of fit.

Let us suppose that $O_1, O_2, ..., O_n$ are the observed (experimental) frequencies and $E_1, E_2, ..., E_n$ are the expected (theoritical) frequencies corresponding to $O_1, O_2, ..., O_n$. if we are interested to test the fitness between the observed and expected frequences we setup the statistical hypothesis as

H_0 : The fit between observed and expected frequencies is good i.,e., there is no significant deviation between observed and expected frequencies.

Vs

H_1 : The fit between observed and expected frequencies is not good i.e., there is a significant deviation between observed and expected frequencies.

The required test statistic to test the above hypothesis is

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right] \sim \chi^2_{(n-1)\text{d.f.}} \text{ at } \alpha \text{ \% l.o.s.}$$

provided $\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i$ under H_0.

If the calculated value of χ^2 is less than the tabulated value of χ^2 for $(n-1)$ d.f. at α% l.o.s., we accept H_0 and conclude that the fit is good between the observed and expected frequencies. Otherwise reject H_0 and conclude that the fit is not good i.e., there is a significant difference between the observed and expected frequencies.

Problem 1: 200 digits were choosen at random from a set of tables. The frequencies of the digits were

Digits :	0	1	2	3	4	5	6	7	8	9
Frequency:	18	19	23	21	16	25	22	20	21	15

Test whether the digits are uniformly distributed in the table.

Solution: Here the hypothesis to be tested is

H_0 : The digits are uniformly distributed

Vs

H_1 : The digits are not uniformly distributed.

Now, the given frequencies are the observed frequencies and we need to calculate the corresponding expected frequencies.

Expected frequencies are calculated by taking the average of the observed frequencies.

Here the average of observed frequencies is $\frac{200}{10}=20$ which is taken as expected frequency corresponding to each observed frequency.

Digits	O_i	E_i	$(O_i - E_i)^2 / E_i$
0	18	20	0.2
1	19	20	0.05
2	23	20	0.45
3	21	20	0.05
4	16	20	0.8
5	25	20	1.25
6	22	20	0.2
7	20	20	0
8	21	20	0.05
9	15	20	1.25
			4.3

$$\chi^2 = \sum_{i=1}^{10}\left(\frac{(O_i - E_i)^2}{E_i}\right) = 4.3 \sim \chi^2_{(10-1)} \text{ d.f. at 5\% l.o.s.}$$

Tabulated value of χ^2 for 9 d.f. at 5% l.o.s. is 16.919

Since, the calculated value of χ^2 is less than the tabulated value of χ^2 for 9 d.f. at 5% l.o.s., we accept H_0 and conclude that the digits are uniformly distributed.

Problem 2: The theory predicts the proportion of beans, in the four groups A, B, C, D should be 9 : 3 : 3 : 1. In an experiment among 1600 beans, the numbers in the four groups were 882, 313, 287 and 118. Does the experimental result support the theory.

Solution: In this problem we set up the hypothesis as

H_0 : The experiment supports the theory.

Vs

H_1 : The experiment does not support the theory.

Under H_0, 1600 benas are distributed in the ratio 9 : 3 : 3 : 1 for A, B, C, D respectively, then the expected frequencies are

$$E(A) = \frac{9}{16}\times 1600 = 900$$

$$E(B) = \frac{3}{16}\times 1600 = 300$$

$$E(C) = \frac{3}{16}\times 1600 = 300$$

$$E(D) = \frac{1}{16}\times 1600 = 100$$

Now,

O_i	E_i	$(O_i - E_i)^2 / E_i$
882	900	0.36
313	300	0.563
287	300	0.563
118	100	3.24
		4.726

$\therefore\ \chi^2 = 4.726 \sim \chi^2_{3\,\text{d.f.}}$ at 5% l.o.s.

Tabulated value of χ^2 for 3 d.f. at 5% l.o.s. is 7.815

Since, the calculated value of χ^2 is less than the tabulated value of χ^2 for 3 d.f. at 5% l.o.s., we accept H_0 and conclude that the experiment supports the theory.

Problem 3: A survey of 320 families with 5 children revealed the following distribution

No. of boys:	5	4	3	2	1	0
No. of girls :	0	1	2	3	4	5
No. of families :	14	56	110	88	40	12

Is the result consistent with the hypothesis that male and female births are equally probable.

Solution: In this problem we setup the hypothesis as

H_0 : Male and female births are equally probable.

Vs

H_1 : Male and female births are not equally probable.

Under H_0, probability of female birth is $1/2$. i.e., $p = \frac{1}{2}$. Let x be the number of female births out of 5 children in each family, then X follows binomial distribution.

$$\therefore\ p(x) = {}^nC_x\, p^x\, q^{n-x}\ ;\ x = 0,1,2,\dots\dots\dots\dots,n$$

and the expected frequency of X is given by $N \cdot p(x)$ i.e., $320 \cdot p(x)$

Now, the expected frequencies are computed taking n = 5, and $x = 0,1,2,3,4,5$.

X	p(x)	Expected freq. $N\,p(x)$
0	0.03125	10
1	0.15625	50
2	0.3125	100
3	0.03125	100
4	0.15625	50
5	0.3125	10

O_i	E_i	$(O_i - E_i)^2 / E_i$
14	10	1.6
56	50	0.72
110	100	1.00
88	100	1.44
40	50	2.00
12	10	0.4
		7.16

$\therefore$ Calculated value of $\chi^2 = 7.16 \sim \chi_4^2$ d.f. at 5% l.o.s.

Tabulated value of χ^2 for 4 d.f. at 5% l.o.s. is 11.07

Since, the calculated value of χ^2 is less than the tabulated value of χ^2, we accept H_0 and conclude that male and female births are equally probable.

14.5 F - test for equality of two population variances:

Let $x_1, x_2, \ldots\ldots\ldots\ldots, x_i, \ldots\ldots\ldots x_{n_1}$ be a random sample from a normal population with unknown variance σ_1^2. Let χ_1^2 be a chi - square variate and S_X^2 be the sample variance defined as

$$\chi_1^2 = \frac{\sum_{i=1}^{n_1}\left(X_i - \overline{X}\right)^2}{\sigma_1^2} \quad \text{and} \quad S_X^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_i - \overline{X}\right)^2$$

Let $y_1, y_2, \cdots\cdots\cdots, y_j, \cdots\cdots\cdots, y_{n_2}$ be a random sample from another normal population with unknown variance σ_2^2. Let χ_2^2 be a chi - square variate and S_Y^2 be the sample variance defined as

$$\chi_2^2 = \frac{\sum_{j=1}^{n_2}\left(Y_j - \overline{Y}\right)^2}{\sigma_2^2} \quad \text{and} \quad S_Y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(y_j - \overline{Y}\right)^2$$

Now, if we are interested to test the equality of the two normal population variances σ_1^2 and σ_2^2, we need to setup the statistical hypothesis as follows:

H_0 : The two normal populations are equal or homogeneous.

i.e., $H_0 : \sigma_1^2 = \sigma_2^2$.

Vs

H_1 : The two normal populations are not equal or heterogeneous.

i.e., $H_1 : \sigma_1^2 \neq \sigma_2^2$.

The required test statistic to test the above hypothesis is

$$F = \frac{\chi_1^2 / n_1 - 1}{\chi_2^2 / n_2 - 1} \sim F(n_1 - 1,\ n_2 - 1)\ \text{d.f. under } H_0 \text{ at } \alpha \text{ \% l.o.s.}$$

$$= \frac{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\frac{X_i - \overline{X}}{\sigma_1} \right)^2}{\frac{1}{n_2 - 1} \sum_{j=1}^{n_2} \left(\frac{Y_j - \overline{Y}}{\sigma_2} \right)^2} \sim F_{(n_1 - 1,\ n_2 - 1)} \quad \text{under } H_0$$

Under $H_0 \quad \sigma_1^2 = \sigma_2^2$,

$$F = \frac{\frac{1}{n_1 - 1} \Sigma \left(X_i - \overline{X} \right)^2}{\frac{1}{n_2 - 1} \Sigma \left(Y_j - \overline{Y} \right)^2} \sim F_{(n_1 - 1,\ n_2 - 1)}$$

$$F = \frac{S_X^2}{S_Y^2} \sim F_{(n_1 - 1,\ n_2 - 1)} \text{ at 5\% l.o.s.}$$

$$\because \quad (n - 1) S^2 = ns^2 \Rightarrow S^2 = \frac{ns^2}{n - 1}, \quad \text{where } s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}{n}$$

$$\therefore \ F = \frac{\frac{n_1 s_x^2}{n_1 - 1}}{\frac{n_2 s_y^2}{n_2 - 1}} \sim F_{(n_1 - 1,\ n_2 - 1)} \ \text{d.f. at 5\% l.o.s.}$$

If the calculated value of F is less than the tabulated value of F for $(n_1 - 1,\ n_2 - 1)$ d.f. at $\alpha\%$ l.o.s., we accept H_0 and conclude that the two normal population variances are equal. Otherwise, we reject H_0 and conclude that the two normal population variances are not equal.

Problem 1: In one sample of 8 observations, the sum of squars of deviations of the sample values from the sample mean was 84.4 and in other sample of 10 observations it was 102.6. Test whether this difference is significant at 5% level, given that the 5% point of F for $n_1 = 7$ and $n_2 = 9$ d.f. is 3.29.

Solution: Here, the hypothesis to be tested is

H_0 : Two normal population variances are equal.

i.e., $H_0 : \sigma_1^2 = \sigma_2^2$.

Vs

H_1 : Two normal population variances are not equal.

i.e., $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Also, we are given

$$n_1 = 8 \quad ; \quad \Sigma\left(X_i - \overline{X}\right)^2 = 84.4$$

$$n_2 = 10 \quad ; \quad \Sigma\left(Y_j - \overline{Y}\right)^2 = 102.6$$

$$\therefore\ S_X^2 = \frac{1}{n_1 - 1}\Sigma\left(X_i - \overline{X}\right)^2 = \frac{1}{7} \times 84.4 = 12.057$$

$$S_Y^2 = \frac{1}{n_2 - 1}\Sigma\left(Y_j - \overline{Y}\right)^2 = \frac{1}{9} \times 102.6 = 11.4$$

Hence, the required test statistic is

$$F = \frac{S_X^2}{S_Y^2} \sim F_{(n_1 - 1,\ n_2 - 1)\ \text{d.f.}} \text{ at } \alpha\% \text{ l.o.s.}$$

$$F = \frac{12.057}{11.4} = 1.057$$

Tabulated value of F for $(7, 9)$ d.f. at 5% l.o.s. is 3.29

Since, the caluculated value of F is less than the tabulated value of F for (7, 9) d.f. at 5% l.o.s., we accept H_0 and conclude that both the population variances are identical.

Problem 2: Pumpkins were grown under two experimental conditions. Two random samples of 11 and 9 pumpkins show the sample standard deviations of their weights as 0.8 and 0.5 respectively. Assuming that the weight distributions are normal, test the hypothesis that the true variances are equal at 5% level of significance.

Solution: Here the hypothesis to be tested is

H_0 : The variances of both the normal populations are equal.

i.e., $H_0 : \sigma_1^2 = \sigma_2^2$.

Vs

H_1 : The variances of both the normal populations are not equal.

i.e., $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Also, we are given

$n_1 = 11$; $s_x = 0.8$

$n_2 = 9$; $s_y = 0.5$

The required test statistic to test the above hypothesis is

$$F = \frac{S_x^2}{S_y^2} \sim F_{(n_1-1,\ n_2-1)\ d.f.} \text{ at } \alpha\% \text{ l.o.s.}$$

$$F = \frac{\dfrac{n_1 s_1^2}{n_1 - 1}}{\dfrac{n_2 s_2^2}{n_2 - 1}} \sim F_{(n_1-1,\ n_2-1) d.f.}$$

$$F = \frac{\dfrac{11 \times (0.8)^2}{11-1}}{\dfrac{9 \times (0.5)^2}{9-1}} \sim F_{(10,\ 8)\ d.f.} \text{ at } 5\% \text{ l.o.s.}$$

$$F = \frac{0.704}{0.28125} = 2.5$$

Tabulated value of F for $(10, 8)$ d.f. at 5% l.o.s. is 3.35

Since, the calculated value of F is less than the tabulated value of F for $(10, 8)$ d.f. at 5% l.o.s., we accept H_0 and conclude that the variances of both the normal populations are equal.

14.6 Exercise:

1. It is belived thatthe precision (as measured by variance) of an instrument is no more than 0.16. Write down the null and alternative hypothesis for testing this belief. Carry out the test at 1% level given 11 measurments of the same subject on the instrument:

   2.5, 2.3, 2.4, 2.3, 2.5, 2.7, 2.5, 2.6, 2.6, 2.7, 2.5

2. Test the hypothesis that $\sigma = 10$, given that, $S = 15$ for a random sample of size 50 from a normal population.

3. A sample of 20 observations gave a standard deviation of 3.72. Is this compatible with the hypothesis that the sample is from a normal population with variance 4.35.

4. A machine puts out defective pins with a S.D. of 3 pins per hour. To test this claim the no of defective pins per 10 hours by the machine are observed as 4,3, 5, 7, 8, 1, 3, 6, 4, 2. Can you accept the claim at 1% level.

5. A manufacturer of gun powder has developed a new powder which is designed to produce muzzle velocity of S.D. equal to 3000 feet/sec. 7 shells are loaded with the charge and they gave the muzzle velocities with a S.D. of 2998.5 feet/sec. Do these data present sufficient evidence to indicate that the S.D. of velocity differs from 3000 feet/sec.

6. The following data is collected on two characters

	Cine Goers	Non-Cinegoers
Literate	83	57
Illeterate	45	68

Based on these can you conclude that there is no relation between habit of cinema going and literacy.

7. For the data in the following table test for independence between a persons ability in mathematics and interest in statistics.

		Ability in maths		
		Low	**Average**	**High**
Interest in Stat	Average	58	61	31
	High	14	47	29

8. The table below shows results of a survey in which 250 respondents were catogirised according to the level ofeducation and attitude towards the students demonstration at a certain college. Test the hypothesis that the two catogiries of classification are independent at 1% l.o.s.

		Attitude		
		Against	**Nuteral**	**for**
Education	Primary	40	25	5
	High school	40	20	5
	Graduate	30	15	30
	Post Graduate	15	15	10

9. In a certain sample of 2,000 families 1400 families are consumers of tea. Out of 1800 Hindu families, 1236 families consume tea. Use χ^2 test and state whether there is any significant difference between consumption of tea among Hindu and non-Hindu families.

10. Out of a sample of 120 persons in a village, 76 persons were administered a new drug for preventing influenza and out of them, 24 persons were attacked by influenza. Out of those who were not administered the new drug, 12 persons were not affected by influenza. Prepare (a) 2×2 table showing actual and expected frequencies

(b) use chi -square test for finding out whether the new drug is effective or not.

11. Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The results are as follows:

Researcher	Below Average	No. of studetns in each level		
		Average	Above Average	Genius
X	86	60	44	10
Y	40	33	25	2

Would you say that the sampling techniques adopted by the two researchers are significantly different. (Given 5% value of χ^2 for 3 d.f. and 4 d.f. are 7.82 and 9.49 respectively.)

12. The demand for a particular spare part in a factory was found to vary from day - to - day. In a sample study the following information was obtained:

Days:	Mon	Tue	Wed	Thurs	Fri	Sat
No.of. Parts demanded:	1124	1125	1110	1120	1126	1115

Test the hypothesis that the number of parts does not depend on the day of the week. (Given: the values of chi - square significance at 5, 6, 7 d.f. are respectively 11.07, 12.59, 14.07 at 5% l.o.s.)

13. The following figures show the distribution of digits in numbers choosen at random from a telephone directory:

Digits:	0	1	2	3	4	5	6	7	8	9
Frequency:	1026	1107	997	966	1075	933	1107	972	964	853

Test whether the digits may be taken to occur equally frequently in the dictionary.

14. A sample analysis of examination results of 200 MBA's was made. It was found that 46 students had failed, 68 secured a third division, 62 secured a second division and the rest were placed in the first division. Are these figures commensurate with the general examination result which is in the ratio of $4:3:2:1$ for various categories respectively.

15. When the first proof of 392 pages of a book of 1200 pages were read, the distribution of printing mistakes were found to be as follows:

No.of. mistakes in a page (x) :	0	1	2	3	4	5	6
No. of. pages (f) :	275	72	30	7	5	2	1

Fit a poisson distribution to the above data and test the goodness of fit.

16. It is known that the mean diameters of rivets produced by two firms A and B are practically the same but the standard deviations may differ. For 22 rivets produced by firm A, the standard deviation is 2.9 mm, while for 16 rivets manufactured by firm B, the standard deviation is 3.8 mm. Compute the statistic you would use to test whether the products of firm A have the same variability as those of firm B and test its significance.

17. Two random samples gave the following results:

Sample	**Size**	**Sample mean**	**Sum of squares of deviations from mean**
1	10	15	90
2	12	14	108

Test whether the two samples come from the same normal population.

18. The following are the values in thousands of an inch obtained by two engineers in 10 successive measurments with the same micrometer. Is one engineer significantly more consistent than the other?

Engineer A:	503, 505, 497, 505, 495, 502, 499, 493, 510, 501
Engineer B:	502, 497, 492, 498, 499, 495, 497, 496, 498

19. The following figures give the prices in rupees of a certain commodity in a sample of 15 shops selected at random from a city A and those in a sample of 13 shops from another city B.

City A :	7.41	7.77	7.44	7.40	7.38	7.93	7.58	
	8.28	7.23	7.52	7.82	7.71	7.84	7.63	7.68
City B :	7.08	7.49	7.42	7.04	6.92	7.22	7.68	
	7.24	7.74	7.81	7.28	7.43	7.47		

Assuming that the distribution of prices in the two cities is normal, answer the following:

(i) Is it possible that the average price of city B is Rs 7.20?

(ii) Is the observed variance in the first sample consistent with the hypothesis that the standard deviation of prices in city A is Rs 0.30?

(iii) Is it reasonable to say that the variability of prices in the two cities is the same?

(iv) Is it reasonable to say that the average prices are same in the two cities.

14.7 Summary:

This lesson in detail discusses about applications of χ^2 and F distributions. χ^2 distribution is given in detail for test of significance of Normal population variance, independence of attributes, goodness of fit. F - distribution is explained for testing the equality of two normal population variances. The exercises given at the end of the lesson makes the student identify different testing situations and the distributions involved for the study.

14.8 Technical Terms:

Test for significance of Normal population variance

Test for independence of attributes

Test for goodness of fit

Test for equality of two Normal population variances

Lesson Writer

P. Nagamani

Lesson 15

NON - PARAMETRIC TESTS

Objective:

After studying this lesson the students will be conversent with the theoritical concepts as well as practical applications of Run test, Sign test and Wilcoxon-Signed Rank test for single sample, sign test and Wilcoxon - signed Rank test for two related samples.

Structure of The Lesson:

15.1 Introduction To Non - Prametric Tests:

Generally parametric tests are based on two features

(i) The parent population and its functional form from which the samples have been drawn is assumed to be known, and

(ii) They are concerned with testing statistical hypothesis about the parameters of this frequency function or estimating its parameters. Therefore, parametric tests are the tests which deal with parameters of the population, from which the sample is drawn.

On the other hand non - parametric tests do not depend on the form of the population from which the sample is drawn. In otherwords non - parametric tests do not make any assumption regarding the form of the population. However, certain assumptions associated with non - parametric tests are

(1) Sample observations are independent

(2) The variable under study is continuous

(3) The p.d.f. should be continuous

(4) Lower order moments exists.

15.2 Advantages and Disadvantages of Non - Parametric Tests:

The following are some of the advantages of N - P tests.

(1) Non - Parametric methods are readily comprehensible, very simple and easy to apply and do not require complicated sampling theory.

(2) No assumption is made about the distribution of parent population from which sampling is done.

(3) N - P methods are applied to deal with data which is measured in nominal.

(4) N - P tests are applied in sociology, Psychometry, economics and educational statistics.

(5) N - P tests are used to deal with data which are given in ranks or grades or scores.

Disadvantages:

Inspite of the above advantages N - P tests have the following disadvantages.

(1) Non - Parametric tests can be used only if the measurments are nominal, even in that case if a parametric test exists, then it is more powerful than N - P test.

(2) These tests are designed to test statistical hypothesis only and not for estimating the parameters.

(3) N - P tests have no applications in analysis of variance for interaction effects.

Some of the tests that are going to be discussed in this lesson are

(1) One sample run test

(2) One sample and two sample sign test

(3) One sample and two sample Wilcoxon Signed Rank test.

15.3 One Sample Run Test:

Run test is based on the runs scored by the sample. A run is defined as a sequence of letters of one kind surrounded by a sequence of letters of other kind.

Test Procedure: Let $x_1, x_2, ..., x_n$ be the set of observations which occur in their natural order from a population. If we want to test whether the sample observations are drawn at random, we construct the hypothesis for testing as

H_0 : The sample observations are drawn at random.

Vs

H_1 : The sample observations are not drawn at random.

The test procedure is to find the Median M, for the given sample. Now indicate the observations by the letter A which are less than M and the observations by the letter B which are greater than M for their natural order. Here we get the sequence of letters $\overline{A}\ \overline{B}\ \overline{AAA}\ \ \overline{BB}\ A$ (say), counting the number of runs 'r'. If $r_1 < r < r_2$ where r_1 and r_2 are the values obtained from the run tables corresponding to $n_1,\ n_2$ at α % l.o.s., where n_1 is number of A's and n_2 is number of B's, we accept H_0 and conclude that the sample observations are drawn at random, otherwise reject H_0 and conclude that the sample observations are not drawn at random. For large samples usually for n_1 or n_2 $> 20,\ r \sim N\left[E(r),\ V(r)\right],$ where

$$E(r) = \frac{2n_1\, n_2}{n_1 + n_2} + 1 \quad \text{and} \quad V(r) = \frac{2n_1 n_2\left(2n_1 n_2 - n_1 - n_2\right)}{\left(n_1 + n_2\right)^2\left(n_1 + n_2 - 1\right)}$$

and the test statistic to test H_0 Vs H_1 is

$$Z = \frac{r - E(r)}{\sqrt{V(r)}} \sim N(0,1).$$

If the calculated value of $|Z|$ is less than the tabulated value of Z at α% l.o.s., we accept H_0 and conclude that the observations are drawn at random, otherwise reject H_0.

Problem 1: Test whether the following observations are drawn at random or not.

109, 124, 173, 167, 148, 132, 168, 165, 118, 112, 114, 164, 180, 123, 180, 152.

Solution: The hypothesis to be tested for the above problem is

H_0 : The sample observations are drawn at random

Vs

H_1 : Sample observations are not drawn at random.

First calculate median. Arrange the observations in asending order.

109, 112, 114, 118, 123, 124, 132, 143, 152, 164, 165, 167, 168, 173, 180, 180.

Now, Median = 147.5

Consider the original (given) data

$$\overline{\overset{A}{109},\ \overset{A}{124}},\ \overline{\overset{B}{173},\ \overset{B}{167},\ \overset{B}{148}},\ \overline{\overset{A}{132}},\ \overline{\overset{B}{168},\ \overset{B}{165}},\ \overline{\overset{A}{118},\ \overset{A}{112},\ \overset{A}{114}},\ \overline{\overset{B}{164},\ \overset{B}{180}},$$

$$\overline{\overset{A}{123}},\ \overline{\overset{B}{150},\ \overset{B}{152}}.$$

The number of runs 'r' = 8.

Also, n_1 = number of A's = 7

n_2 = number of B's = 9

Limits for $n_1 = 7$ and $n_2 = 9$ at 5% l.o.s. using run tables (see prof. T.V. Avadhani "Statistical Tables" p. 34-35) are $r_1 = 4$; $r_2 = 14$ i.e., 4 < 8 < 14

Since $r_1 < r < r_2$ we accept H_0 and conclude that the sample observations are drawn at random.

Problem 2: In a dinner party men and women sat in a row in the following order

MWMWMMMWMWWWWMWWMWMWMMWMWWWMWMWM

Test whether they sat at random.

Solution: The hypothesis to study the above problem is

H_0 : Men and Women sat at random

Vs

H_1 : Men and Women did not sit at random

Now, consider the data for getting runs

$$\overline{M}\,\overline{W}\,\overline{M}\,\overline{W}\,\overline{MMM}\,\overline{W}\,\overline{M}\,\overline{WWWW}\,\overline{M}\,\overline{WW}\,\overline{M}\,\overline{W}\,\overline{M}\,\overline{W}\,\overline{MM}\,\overline{W}\,\overline{M}\,\overline{WWW}\,\overline{M}\,\overline{W}\,\overline{M}\,\overline{W}\,\overline{M}$$

Here r = 23

n_1 = number of Men = 15

n_2 = number of Women = 17

The values of r_1 and r_2 corresponsing to n_1, n_2 at $\alpha = 0.05$ l.o.s. from run tables are

$r_1 = 11$; $r_2 = 23$

since $r = r_2$, we reject H_0 and conclude that the sitting pattern of men and women are not at random.

Problem 3: A coin is tossed 37 times and head occurred 25 times and their number of runs are 13. Test whether the coin is unbiassed.

Solution: The hypothesis to be tested is

H_0 : The coin is unbiassed

Vs

H_1 : The coin is not unbiassed

Given number of runs r = 13.

Number of heads n_1 = 25

Number of tails n_2 = 12

$\because n_1 > 20$, we consider as large sample

The required statistic to test is

$$Z = \frac{r - E(r)}{\sqrt{V(r)}} \sim N(0,1)$$

$$E(r) = \frac{2n_1n_2}{n_1+n_2} + 1 = \frac{2 \times 25 \times 12}{25+12} + 1$$

= 17.2162

$$V(r) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2 (n_1+n_2-1)}$$

$$= \frac{2 \times 25 \times 12(2 \times 25 \times 12 - 25 - 12)}{(25+12)^2 (25+12-1)}$$

$$V(r) = 6.8542$$

$$\therefore Z = \frac{r - E(r)}{\sqrt{V(r)}} = \frac{13 - 17.2162}{2.6180} = -1.6104$$

Tabulated value of $|Z|$ at 5% l.o.s. is 1.96

Since calculated value of $|Z|$ is less than the tabulated value of Z at 5% l.o.s., we accept H_0 and conclude that the coin is unbiased.

15.4 Sign Test:

Sign test is the simplest of all the non - parametric tests. As the name comes sign test is based on the direction of change of observations but not on their numerical magnitude. Sign test is used to study whether the sample has come from the population with a specified median. Sign test can be applied for single sample as well as two related samples (paired data) which are discussed in detail.

15.5 Sign test for one sample:

Let $x_1, x_2, ..., x_n$ be a random sample drawn from an unknown continuous population with median M. If we want to test the population median $M = M_0$, we set up the statistical hypothesis as follows:

H_0 : The sample has been drawn from the population with median $M = M_0$

Vs

H_1 : The sample has not been drawn from the population with median $M \neq M_0$.

The test procedure to test the hypothesis is first arrange the sample observations in the ascending order. i.e., $x_1 < x_2 < .... < x_n$. Let $d_i = x_i - M_0$ under H_0. Consider the signs of deviations d_i excluding 0 deviations. we have $P_r\{d_i < 0\} = P_r\{d_i > 0\} = 1/2$. [$\because$ M_0 is the median which divides the entire data into two equal parts]. Let 'x' be number of +ve or -ve signs of d_i out of n signs which follows binomial with parameters n, $p = 1/2$, then

$$p(x) = {}^nC_x\, p^x\, (1-p)^{n-x}.$$

$$= {}^nC_x \left(\frac{1}{2}\right)^x \left(1-\frac{1}{2}\right)^{n-x}$$

$$p(x) = {}^nC_x \left(1/2\right)^n$$

The test criteria is to compute $2F_x(u) = 2P(X \leq u)$

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

$$= 2 \cdot \sum_{X=0}^{u} {}^{n}C_{X} \left(\frac{1}{2}\right)^{n}$$

$$= \frac{1}{2^{n-1}} \sum_{X=0}^{u} {}^{n}C_{X}$$

where u is number of +ve or -ve signs which ever is least.

If the calculated value of $2F_X(u) > \alpha$, we accept H_0 and conclude that the given sample has been drawn from the population with median $M = M_0$. Otherwise, we reject H_0 and conclude that the sample has not been drawn from the population with median $M = M_0$. Particularly, if the sample is large i.e., $n \geq 25$, we use standard normal test criteria to test above H_0.

Since, $X \sim B\left(n, p = \frac{1}{2}\right)$

$$Z = \frac{X - E(X)}{S \cdot E \cdot (X)} = \frac{X - n/2}{\sqrt{n/4}} = \frac{u - n/2}{\sqrt{n}/2}$$

$$\therefore \quad Z = \frac{2u - n}{\sqrt{n}} \sim N(0,1) \text{ under } H_0$$

If the calculated value of $|Z|$ < tabulated value of Z at $\alpha\%$ l.o.s., we accept H_0, otherwise reject H_0.

Problem 1: A random sample of 20 observations drawn from a populataion are 93, 88, 107, 115, 82, 97, 103, 86, 113, 107, 112, 90, 98, 93, 99, 103, 100, 101, 96, 104. Test the population median is 90 at 5% level of significance.

Solution: Here the hypothesis to be tested is

H_0 : The population median is 90

i.e., H_0 : $M = 90$

Vs

H_1 : The population median is not 90

i.e., H_1 : $M \neq 90$

Let us first arrange the data in ascending order.

82, 86, 88, 90, 93, 93, 96, 97, 98, 99, 100, 101, 103, 103, 104, 107, 107, 112, 113, 115.

Now we compute

$$d_i = X_i - M = X_i - 90$$

$$d_i = -8, -4, -2, 0, 3, 3, 6, 7, 8, 9, 10, 11, 13, 13, 14, 17, 17, 22, 23, 25$$

Since there is one zero deviation, the size of the sample n = 19.

Total number of +ve signs = 16

Total number of -ve signs = 3

$$u = \min(+ve, -ve) = \min(16, 3) = 3$$

The test critiria is $2F = \frac{1}{2^{n-1}} \sum_{X=0}^{u} {}^{n}C_X$

$$= \frac{1}{2^{18}} \left[\sum_{X=0}^{3} {}^{19}C_X \right]$$

$$= \frac{1}{2^{18}} \left[{}^{19}C_0 + {}^{19}C_1 + {}^{19}C_2 + {}^{19}C_3 \right]$$

$$= \frac{1160}{262144}$$

$$= 0.0044$$

The critical value is 0.05

Since, the calculated value is less than the critical value, we reject H_0 and conclude that the population median is not 90.

Problem 2: In a certain college the students are classified according to their weights below 60 kgs and above 60 kgs. A random sample of 100 students is takes from the college out of which 30 students belong to I classification. Based on this information test that the two catogories of the classification of the students ae in equal ratio in the college.

Solution: Let I classification represents the students < 60 kgs weight.

Let II classification represents the students > 60 kgs weight.

Here the hypothesis to be tested is

H_0 : The two classifications are equal in ratio.

i.e., $H_0 : M = 60$

Vs

H_1 : The two classifications are not equal in ratio.

i.e., $H_1 : M \neq 60$.

Also, given that n = 100

Students of Ist classification is 30. i.e., number of

- ve signs = 30 and

+ ve signs = 100 - 30 = 70

$u = \min(30,\ 70) = 30$

$\because\ n \geq 25$ i.e., large, the test criteria is

$$Z = \frac{2u - n}{\sqrt{n}} = \frac{2 \times 30 - 100}{\sqrt{100}} = -4$$

$$|Z| = 4$$

Table value of $|Z|$ at 5% l.o.s. is 1.96

Since, the calculated value of $|Z| >$ tabulated value of Z at 5% l.o.s., we reject H_0 and conclude that the two classifications are not in the equal ratio.

Problem 3: A sample of diameters (inches) of 10 individual parts produced from a certain industry are 54, 51, 61, 63, 68, 62, 56, 58, 53, 55. Test the median diameter is 61.

solution: The hypothesis to betested is

H_0 : The median diameter is 61 inches.

i.e., $H_0 : M = 61$ inches.

Vs

H_1 : The median diameter is not 61 inches.

i.e., $H_1 : M \neq 61$ inches.

Let us now order the data

51, 53, 54, 55, 56, 58, 61, 62, 63, 68

Given M = 61

Now $d_i = X_i - M = X_i - 61$.

$$d_i = -\text{ve}, -\text{ve}, -\text{ve}, -\text{ve}, -\text{ve}, -\text{ve}, 0, +\text{ve}, +\text{ve}, +\text{ve},$$

Since there is one zero deviation, the sample size n = 9

Total number of +ve's = 3

Total number of -ve's = 6

$$u = \min(3,6) = 3$$

The test criteria to test the above H_0 is

$$2F = \frac{1}{2^{n-1}} \sum_{X=0}^{u} {}^{n}C_X \sim \alpha\% \ \ell.\text{o.s}$$

$$2F = \frac{1}{2^{9-1}} \left[\sum_{X=0}^{3} {}^{9}C_X \right]$$

$$= \frac{1}{2^8} \left[{}^{9}C_0 + {}^{9}C_1 + {}^{9}C_2 + {}^{9}C_3 \right]$$

= 0.0039 (130)

= 0.507

Table value is 5% = 0.05

Since the caluculated value is greater than the tabulated value we accept H_0 and conclude that the median diameter is 61 inches.

15.6 Sign test for two related samples:

Let $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ be a sample of size 'n' from an unknown continuous population with its form $f(x)$.

Let $y_1, y_2, \cdots\cdots\cdots\cdots, y_n$ be another random sample of size 'n' from the other unknown continuous population with its form $f(y)$.

If we want to test the significant difference between the two populations, the hypothesis for testing is

H_0 : There is no significant difference between the two populations

i.e., $H_0 : f(x) = f(y)$

Vs

H_1 : There is a significant difference between the two populations.

i.e., $H_1 : f(x) \neq f(y)$

The test procedure is to observe the signs of deviations $d_i = (x_i - y_i)$. Now exclude '0' deviations and take 'U' as number of +ve signs or -ve signs which ever is least.

Now, get the values of u_1 and u_2 at $\alpha/2$ and $1-\alpha/2$ l.o.s. respectively for the sample size 'n' obtained from binomial cumulative probabilities.

If $u_1 < u < u_2$, we accept H_0 and conclude that there is no significant difference between the two populations. Otherwise reject H_0 and say taht there is a significant difference between the two populations.

Particularly, for $n \geq 25$ (large), then the test criteria is

$Z = \dfrac{2u - n}{\sqrt{n}} \sim N(0,1)$ under H_0, to test the above hypothesis and if caluculated $|Z|$ is less than the critical value of Z we accept H_0, otherwise reject H_0.

Problem 1: 17 students were selected for special coaching, the following data reveals the grades of the students in the examinations before and after they were given coaching. Test whether the performances improved after the coaching.

Grades before coaching :	2 3 3 3 3 3 3 3 2 3 2 2 6 2 5 3 1
Grades after coaching :	4 4 5 5 3 2 5 3 1 5 5 5 4 5 5 5 5

Solution: The hypothesis to be tested is

H_0 : The performance of students in an examination is same before and after coaching.

i.e., $H_0 : f(x) = f(y)$.

Vs

H_1 : The performance of students in an examination is not same before and after coaching.

i.e., $H_1 : f(x) \neq f(y)$.

x_i	:	2	3	3	3	3	3	3	3	2	3	2	2	6	2	5	3	1
y_i	:	4	4	5	5	3	2	5	3	1	5	5	5	4	5	5	5	5
$d_i = (x_i - y_i)$	:	-	-	-	-	0	+	-	0	+	-	-	-	+	-	0	-	-

Here, the sample size is 17.

As we have 3 zero deviations, the size of the sample reduces by 3.

i.e., n = 17 - 3 = 14.

If $\alpha = 5\%$, the value of u_1 at $\alpha/2 = 0.025$ for n = 14 is 2.

The value of u_2 at $1 - \alpha/2 = 1 - 0.025 = 0.975$ for n = 14 is 11.

No of +ve's = 3 ; No of - ve's = 11

Here $u_1 < u < u_2; 2 < 3 < 11$, hence we accept H_0 and conclude that the performance of the students in an examination is same i.e., it is not improved even after the coaching.

Problem 2: In a firm 30 machines were selected at random and observed the number of defectives produced by them before and after servicing the machines. Can we conclude that the conditions of machines are improved after their servicing based on the following information

Defectives before servicing	:	133	146	136	172	141	106	159	141	142	140
		175	156	82	138	154	163	175	150	134	134
		160	133	249	129	140	162	115	104	121	181

Defectives after servicing	:	141	151	99	145	179	161	168	151	132	180
		163	133	138	180	204	157	215	145	168	137
		136	170	157	168	188	188	116	146	123	177

Solution: The hypothesis to be tested is

H_0 : There is no change in the condition of machines before and after servicing.

i.e., $H_0 : f(x) = f(y)$.

Vs

H_1 : There is improvement in the conditions of the machines after servicing.

i.e., $H_1 : f(x) > f(y)$

Considering the given data let us first calculate $d_i = (X_i - Y_i)$

$d_i =$	-	-	+	+	-	-	-	-	+	-
	+	+	-	-	-	+	+	+	-	-
	+	-	+	-	-	-	-	-	-	+

n = 30, no of +ves = 10

Since, $n \geq 25$, the test criteria is

$$Z = \frac{2u - n}{\sqrt{n}} \sim Z(0,1)$$

$$= \frac{2(10) - 30}{\sqrt{30}}$$

$$Z = -1.8$$

Table value of Z at 5% l.o.s. one tailed test is 1.645.

Sincethe calculated $|Z|$ is greater than the critical value at 5% l.o.s., we reject H_0 and conclude that there is improvement in the conditions of the machines after servicing.

15.7 Wilcoxon Signed Rank test for single sample:

The sign test was based on only the sign of deviations of the sample values from the hypothesised median M. No attention was paid to the magnitude of the differences. Wilcoxon signed rank test utilises the signs as well as the magnitudes of the differences. This test is more powerful and sensitive than sign test.

Let $x_1, x_2, ..., x_n$ be a random sample drawn from an unknown continuous population with median M. If we are interested to test the population median $M = M_0$, we set up the statistical hypothesis as follows:

H_0 : The given sample has been drawn from the population with the median $M = M_0$.

Vs

H_1 : The given sample has not been drawn from the population with the median $M \neq M_0$.

The test procedure to test the above hypothesis is to first arrange the data in ascending order. Then calculate $d_i = (X_i - M)$ which may have both +ve's and -ve's. Descarding the sign of d_i assign ranks to $d_i \; \forall \; i = 1, ..., n$.

Let t^+ represents the sum of ranks of d_i corresponding to +ve signs and t^- represents the sum of ranks of d_i corresponding to -ve signs. Now, the test criteria is $t = \text{Min}\left(t^+, \; t^-\right)$.

If $t > t_\alpha$, we accept H_0 and conclude that the sample has been drawn from the population with the median $M = M_0$. Here t_α is the critical value obtained from Wilcoxon signed ranks table corresponsing to the sample size 'n' at $\alpha\%$ l.o.s.

Particulary for large samples $n \geq 25, \; t \sim N\left[E(t), V(t)\right]$

$$E(t) = \frac{n(n+1)}{4} \; ; \; V(t) = \frac{n(n+1)(2n+1)}{24}$$

The test criteria is

$$Z = \frac{t - E(t)}{S.E.(t)} = \frac{t - \left(\frac{n(n+1)}{4}\right)}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N\,(0,1)$$

If the calculated value of $|Z|$ is less than the tabulated value of Z at α % l.o.s., we accept H_0, otherwise reject H_0.

Problem 1: A random sample of 20 observations drawn from a population are 93, 88, 107, 115, 82, 97, 103, 86, 113, 107, 112, 90, 98, 93, 99, 103, 100, 101, 96, 104. Test the population median is 90 using Wilcoxon signed rank test.

Solution: The hypothesis to be tested is

H_0 : The population median is 90.

i.e., $H_0 : M = 90$

Vs

H_1 : The population median is not 90.

i.e., $H_1 : M \neq 90$

Let us first arrange the data in ascending order

82, 86, 88, 90, 93, 93, 96, 97, 98, 99, 100, 101, 103, 103, 104, 107, 107, 112, 113, 115

$$d_i = (X_i - M_0)$$

d_i : -8, -4, -2, 0, 3, 3, 6, 7, 8, 9, 10, 11, 13, 13, 14, 17, 17, 22, 23, 25

Rank of $|d_i|$: 7.5, 4, 1, -2.5, 2.5, 5, 6, 7.5, 9, 10, 11, 12.5, 12.5, 14, 15, 16, 17, 18, 19

t^+ is sum of ranks of +ve d_i's = 177.5

t^- is sum of ranks of -ve d_i's = 12.5

$$t = \min\left(t^+,\ t^-\right) = \min(177.5,\ 12.5) = 12.5$$

As one of $d_i = 0, \quad n = 20 - 1 = 19$

Critical value of t_α at 5% l.o.s. for n = 19 from Wilcoxon signed rank tables is 46.

Since $t < t_\alpha$ i.e., 12.5 < 46, we reject H_0 and conclude that the population median is not 90.

Problem 2: A random sample of 30 observations were drawn from the population for studing the population median is 100. It was observed that the sum of ranks of $|d_i|$ w.r.t. +ve signs is 275 and sum of ranks of $|d_i|$ w.r.t -ve signs is 35. Use Wilcoxon signed rank test at 5% l.o.s.

Solution: The hypothesis to be tested is

H_0 : The population median is 100.

i.e., $H_0 : M = 100$

Vs

H_1 : The population median is not 100.

i.e., $H_1 : M \neq 100$

Given that $n = 30; \quad t^{+} = 275; \quad t^{-} = 35$

$$t = \min\left(t^{+},\ t^{-}\right) = 35$$

As $n > 25$, use large sample test.

The test criteria is

$$Z = \frac{t - E(t)}{S.E.(t)}$$

$$E(t) = \frac{n(n+1)}{4} = \frac{30(30+1)}{4} = 232.5$$

$$V(t) = \frac{n(n+1)(2n+1)}{24} = \frac{30(30+1)(2\times 30+1)}{24} = 2363.75$$

$$Z = \frac{35 - 232.5}{48.6184} = -4.062$$

$$|Z| = 4.062$$

Table value of Z at 5% l.o.s. for two tailed test is 1.96.

Since calculated $|Z|$ greater than the tabulated Z at 5% l.o.s., we reject H_0 and conclude that the population median is not 100.

15.8 Wilcoxon - Signed - Rank Test For Two Samples:

Let $x_1, x_2, ..., x_n$ be a random sample drawn from continuous population with its form $f(x)$.

Let $y_1, y_2, \cdots\cdots\cdots\cdots, y_n$ be another random sample drawn from another continuous population with its form $f(y)$.

If we are interested to test the significant difference between the two populations we set up the statistical hypothesis as

H_0 : There is no significant difference between the two populations.

i.e., $H_0 : f(x) = f(y)$

Vs

H_1 : There is a significant difference between the two populations.

i.e., $H_1 : f(x) \neq f(y)$

The test procedure is to give ranks for $|d_i|$, where $d_i = X_i - Y_i \ \forall \ i = 1, 2, ..., n$. Let t^+ be the sum of ranks of $|d_i|$ corresponding to +ve signs and let t^- be the sum of ranks of $|d_i|$ corresponding to $-\text{ve}$ signs. The test criteria is $t = \min\left(t^+, t^-\right)$.

If $t > t_\alpha$ we accept H_0 and conclude that the two populations are identical, otherwise reject H_0, where t_α is the critical value of t at sample size 'n' for $\alpha\%$ l.o.s., obtained from Wilcoxon signed rank tables.

Particulary, for large sample size $n \geq 25$, $t \sim N\left[E(t), V(t)\right]$

Where $E(t) = \dfrac{n(n+1)}{4}$ and $V(t) = \dfrac{n(n+1)(2n+1)}{24}$

Test criteria is $Z = \dfrac{t - E(t)}{S.E.(t)} \sim N(0,1)$

If the calculated value of $|Z| <$ the tabulated value of Z at $\alpha\%$ l.o.s., we accept H_0.

Problem 1: The following are the two samples drawn from two continuous populations. Test whether the two populations have the same distribution function at 5% l.o.s. using two sample Wilcoxon signed rank test.

Sample 1:	7.6	6.3	10.3	6.2	5.4	9.3	10	8.4
Sample 2:	7.3	5.7	10.5	4.7	5.3	8.9	9.1	7

Solution: The hypothesis to betested is

H_0 : The two distribution functions are equal

Vs

H_1 : The two distribution functions are not equal.

Sample I (x_i)	**Sample II** (y_i)	d_i	$\lvert d_i \rvert$	Ranks $\lvert d_i \rvert$
7.6	7.3	0.3	0.3	3
6.3	5.7	0.6	0.6	5
10.3	10.5	-0.2	0.2	2
6.2	4.7	1.5	1.5	8
5.4	5.3	0.1	0.1	1
9.3	8.9	0.4	0.4	4
10	9.1	0.8	0.8	6
8.4	7.0	1.4	1.4	7

t^+ = Sum of ranks of $\lvert d_i \rvert$ corresponding to +ve sign.

$t^+ = 34$

t^- is the sum of ranks of $\lvert d_i \rvert$ corresponding to -ve signs.

$t^- = 2$

$t = \min\left(t^+,\ t^-\right) = \min\left(34, 2\right) = 2$

Critical value of t for sample size n = 8 at α % l.o.s. is 4.

Since $t < t_\alpha$ i.e., 2 < 4, we reject H_0 at 5% l.o.s. and conclude that the two distribution functions are not equal.

Problem 2: The following are the scores of IQ's of 12 pairs of twin children born to 12 parents.

IQ Scroe of CHILD 1 (X) :	86	79	77	68	91	72	77	91	70	77	88	87
IQ Score of CHILD 2 (Y) :	88	77	76	64	96	72	65	90	65	80	81	72

Test whether the child born first has more IQ than second at 1% l.o.s. using Wilcoxon signed rank test.

Solution: The hypothesis for testing the above is

H_0 : The IQ scores are equal for both the children

Vs

H_1 : The chil born first has more IQ than second

X	:	86	79	77	68	91	72	77	91	70	77	88	87
Y	:	88	77	76	64	96	72	65	90	65	80	81	72
d_i	:	-2	2	1	4	-5	0	12	1	5	-3	7	15
$\|d_i\|$	:	2	2	1	4	5	0	12	1	5	3	7	15
R_i	:	3.5	3.5	1.5	6	7.5	-	10	1.5	7.5	5	9	11

t^+ = sum of ranks of $|d_i|$ corresponding to $+\text{ve}$ sign.

$t^+ = 50$

t^- = sum of ranks of $|d_i|$ corresponding to -ve sign.

$t^- = 16$

$$t = \min\left(t^+,\ t^-\right) = \min(50,\ 16) = 16$$

Table value of t for sample size 11 at 1% l.o.s. for single tail is 7.

Since $t > t_\alpha$ i.e., 16 > 7, we accept H_0 and conclude that the IQ scores are equal for both the children.

15.9 Exercises:

1. In a toy shop a particular arrangement of cars (c) and jeeps (j) was displayed in a shelf. Test the randomness of the arrangement

   C C J J J J C C C J J C J C J C C C C C J J J C J J C J C J J J J C C C J C J J J J C C C C J

2. On lossing a coin 19 times, the following sequence of heads (H) and tails (T) was obtained.

   T T H H H T H T H H H T T H H T T H T

   Test whether the coin is unbiased by Run test.

3. Fertilizer X is applied on a particular plant for 35 times and observed that Fertilizer X is successful in improving the yield for 24 times, the number of runs observed were 16. Then test whether the fertilizer is unbiased or not (i.e. is the fertilizer upto the specifications or not).

4. Given the following observations, test the ransomness of the observations.

   246, 224, 232, 184, 146, 192, 252, 199, 272, 240, 210, 212, 215, 186, 192, 199, 248, 172, 136, 148, 232.

5. A typing school claims that in a six - week intensive course, it can train students to type, on the average atleast 60 words per minute. A random sample of 15 graduates is given a typing test and the median number of words per minute typed by each of these students is given below. Test the hypothesis that the median typing speed of graduates is at least 60 words per minute.

   81, 76, 53, 71, 66, 59, 88, 73, 80, 66, 58, 70, 60, 56, 55.

6. Following is the data arranged in order of magnitude of a random sample from a continuous population with median M = 13. Test the data using sign test at 1% l.o.s. 9.62, 10.19, 10.27, 11.13, 12.26, 13.55, 14.16, 14.46, 15.02, 15.92, 16.26.

7 The following 30 numbers are taken from a two - digited number table.

51	68	30	81	90	46	99	98	11	06
19	43	95	82	65	85	65	81	00	50
53	69	51	97	79	69	60	15	05	35

Test the randomness of numbers by the Run test on the basis of runs up and runs down.

8. In a college a random sample of 28 students were taken and their marks in prefinal and final exams in a subject are as follows. Test whether the prefinal and final exam marks have the same distribution function, using Wilconxon - Signed Rank Test.

Prefinal :	183, 175, 134, 170, 183, 167, 120, 175, 126, 187, 123, 121, 175, 133, 144, 109, 165, 144, 164, 125, 183, 175, 134, 170, 183, 167, 120, 175
Final :	133, 198, 170, 164, 199, 160, 168, 158, 162, 176, 126, 141, 103, 126, 126, 155, 162, 161, 182, 119, 133, 198, 170, 164, 199, 160, 168, 158.

9. Use sign test to see if there is a difference between the number of days until the collection of an account receivable before and after a new collection policy. Use 0.05 level of significance.

Before :	30	28	34	35	40	42	33	38	34
After :	32	29	33	32	37	43	40	41	37
Before :	45	28	27	25	41	36			
After :	44	27	33	30	38	36			

15.10 Summary:

This lesson presents in detail about non - parameteric tests, their advantages and disadvantages. Sign test and Wilcoxon - signed rank test are discussed for single sample and two related samples. Also the onesample run test is discussed. A number of problems are solved to explain the above tests and a number of exercises are given to students to solve on their own.

15.11 Technical Terms:

Non - Parametric Tests

Run Test

Sign Test

Wilcoxon - Signed Rank Test

Lesson Writer

P. Nagamani

Lesson 16

NON - PARAMETRIC TESTS FOR TWO INDEPENDENT SAMPLES

Objective:

After studying the lesson the students will have clear comprehension of the theory and practical utility of Median Test, Wicoxon - Mann - Whitney U test and Wald - Wolfowitz Run Test for two independent samples.

Structure of The Lesson:

16.1 Introduction:

In the previous lesson we have discussed some non - parametric tests for one sample and two related samples. In particular, two sample non - parametric tests like sign test and Wilcoxon - Signed Rank tests are carried out for two related samples. In this lesson we see some other non - parametric tests where the two samples involved for the study are independent.

Here, we discuss N - P tests like Median Test, Wilcoxon - Mann - Whitney U test and Wald - Wolfowitz Run Test.

16.2 Wald - Wolfowitz Run Test:

This test is used for testing the equality of two populations. To avoid confusion between one sample run test and this test. We consider two independent samples for study here.

Let $x_1, x_2, ..., x_{n_1}$ be a random sample of size n_1 drawn from unknown continuous population with the density $f(x)$ and let $y_1, y_2, \cdots\cdots\cdots\cdots\cdots, y_{n_2}$ be a sample of size n_2 drawn from another unknown continuous population with density $f(y)$.

If we are interested to test the equality of two populations or two populations with the same density functions, we set up the statistical hypothesis as

H_0 : Two populations are equal.

i.e., $H_0 : f(x) = f(y)$

Vs

H_1 : The two populations are not equal.

i.e., $H_1 : f(x) \neq f(y)$

The test procedure is to combine the two samples as to make a single sample of size $n = n_1 + n_2$ and then arrange the observations in ascending order, then represent the observations by letter A if it belongs to Sample I and the observation by letter B if it belongs to Sample II, for the ascending order of the sample of size 'n'. Then we get the sequence of two letters A and B, now count the number of runs 'r'.

If $r_1 < r < r_2$, where r_1 and r_2 are read from run tables for $n_1,\ n_2$ at 5% l.o.s., we accept H_0 and conclude that there is no significant difference between the two populations, otherwise reject H_0, and conclude that there is a significant difference between the two populations.

For large sample n_1 or n_2 greater than 20, the test statistic is

$$Z = \frac{r - E(r)}{S.E.(r)} \sim N(0,1)$$

where $E(r) = \dfrac{2n_1n_2}{n_1 + n_2} + 1$

and $V(r) = \dfrac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$

If calculated Z < tabulated Z at α % l.o.s., we accept H_0 and conclude that there is no significant difference between the two populations, otherwise we reject H_0.

Problem 1: A Psycho analyst study a particular group of 10 men and 10 Women on a single character and obtained the following scores.

Men	**:**	74	72	77	76	76	73	75	73	74	75
Women	**:**	75	77	78	79	77	73	78	79	78	80

Test whether the men and women are equal in ratio of possessing the character using Run Test.

Solution: The hypothesis to be tested for the above problem is

H_0 : Possessing the character by men and women are equal in ratio.

Vs

H_1 : Possessing the character by men and women are not equal ratio.

First arrange the data in ascending order considering both the samples at a time.

$$\overline{\begin{matrix}A & A & A\\ 72 & 73 & 73'\end{matrix}}\;\overline{\begin{matrix}B\\ 73'\end{matrix}}\quad\overline{\begin{matrix}A & A & A & A\\ 74 & 74 & 75 & 75'\end{matrix}}\quad\overline{\begin{matrix}B\\ 75'\end{matrix}}\quad\overline{\begin{matrix}A & A & A\\ 76 & 76 & 77'\end{matrix}}$$

$$\overline{\begin{matrix}B & B\\ 77 & 77'\end{matrix}}\quad\overline{\begin{matrix}B & B & B & B & B & B\\ 78 & 78 & 78 & 79 & 79 & 80\end{matrix}}$$

A stands for men

B stands for women

Now, number of runs $r=6$.

$n_1=10; \quad n_2=10$

The limits of r at $n_1=10$ and $n_2=10$ at 5% l.o.s. are

$r_1=6$ and $r_2=16$

Since $r=r_1$; i.e., $r_1 \le r < r_2$, we reject H_0, we conclude that the men and women are not in equal ratio in possessing the character.

Problem 2: The following is the data related to rainfall in cms in two places.

Place I	**:**	13, 12, 12, 10, 10, 10, 9, 8, 8, 7, 7, 7, 7, 6, 5, 9, 12, 12, 12, 13, 12.
Place II	**:**	17, 16, 15, 15, 15, 14, 14, 14, 13, 13, 13, 12, 12, 12, 11, 11, 10, 10, 8, 8.

Test whether the two places have the same rain fall.

Solution: The hypothesis to be tested for the above problem is

H_0 : The two places have the same rainfall.

Vs

H_1 : The two places do not have the same rain fall.

Let us first combine the two samples and arrange them in ascending order.

$$\begin{matrix} \overline{A\quad A\quad A\quad A\quad A\quad A\quad A\quad A} & \overline{B\quad B} & \overline{A\quad A\quad A\quad A\quad A} & \overline{B\quad B\quad B\quad B} \\ 5,\ 6,\ 7,\ 7,\ 7,\ 7,\ 8,\ 8, & 8,\ 8, & 9,\ 9,\ 10,\ 10,\ 10, & 10,\ 10,\ 11,\ 11, \end{matrix}$$

$$\begin{matrix} \overline{A\quad A\quad A\quad A\quad A\quad A} & \overline{B\quad B\quad B\quad B} & \overline{A\quad A} \\ 12,\ 12,\ 12,\ 12,\ 12,\ 12, & 12,\ 12,\ 12,\ 12, & 13,\ 13, \end{matrix}$$

$$\begin{matrix} \overline{B\quad B\quad B\quad B\quad B\quad B\quad B\quad B\quad B\quad B\quad B} \\ 13,\ 13,\ 13,\ 14,\ 14,\ 14,\ 15,\ 15,\ 15,\ 16,\ 17. \end{matrix}$$

A Stands for place I

B Stands for place II

$$r = 8$$

$\because n_1,\ n_2 > 20$, we consider the large sample test.

The test criteria is

$$Z = \frac{r - E(r)}{S.E.(r)} \sim N(0,1) \text{ at } \alpha\% \text{ l.o.s.}$$

$$E(r) = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$= \frac{2 \times 21 \times 21}{21 + 21} + 1$$

$$= 22$$

$$V(r) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

$$= \frac{2\times 21\times 21\left(2\times 21\times 21-21-21\right)}{\left(21+21\right)^2\left(21+21-1\right)}$$

$$= 10.2439$$

$$\text{S.E.}(r) = 3.2006$$

$$Z = \frac{8-22}{3.2006}$$

$$|Z| = 4.3741$$

Tabulated value of Z at 5% l.o.s. is 1.96.

Since, calculated $|Z|$ is greater than the tabulated Z at 5% l.o.s., we reject H_0 and conclude that the rainfall in two places is not same.

Problem 3: Test whether the two samples have been drawn at random from a single population.

Sample I :	227, 176, 252, 149, 16, 55, 243, 194, 247, 92, 184, 147, 88, 161, 171.
Sample II :	202, 141, 165, 171, 292, 271, 151, 235, 147, 99, 63, 284, 53, 228, 271.

Solution: The hypothesis to be tested for the above problem is

H_0 : The two samples have been drawn at random from a single population.

Vs

H_1 : The two samples are not drawn at random from a single population.

Let us combine both the samples and arrange in the ascending order.

$$\overline{A}\quad \overline{B}\quad \overline{A}\quad \overline{B}\quad \overline{A\quad A}\quad \overline{B\quad B}\quad \overline{A}\quad \overline{B}\quad \overline{A}\quad \overline{B}\quad \overline{A}\quad \overline{B}\quad \overline{A}\quad \overline{B}$$

$$16,\ 53,\ 55,\ 63,\ 88,\ 92,\ 99,\ 141,\ 147,\ 147,\ 149,\ 151,\ 161,\ 165,\ 171,\ 171,$$

$$\overline{A\quad A\quad A}\quad \overline{B}\quad \overline{A}\quad \overline{B\quad B}\quad \overline{A\quad A\quad A}\quad \overline{B\quad B\quad B\quad B}$$

$$176,\ 184,\ 194,\ 202,\ 227,\ 228\ \ 235,\ 243,\ 247,\ 252,\ 271,\ 271,\ 284,\ 292.$$

A Stands for sample I

B Stands for sample II

Number of runs, $r = 20$

Limits of r_1 and r_2 for $n_1 = 15, \; n_2 = 15$ at 5% l.o.s. are 10 and 22 respectively.

Since, $r_1 < r < r_2$ i.e., 10 < 20 < 22 we accept H_0 and conclude that the two smples have been drawn from a single population.

16.3 Wilcoxon - Mann - Whitney U Test:

This test is described by Wilcoxon and studied by Mann and Whitney. It is most widely used as an alternative to t - test, when the assumptions of t -test does not hold on parent population.

Let $x_1, x_2, ..., x_{n_1}$ and $y_1, y_2, \cdots\cdots\cdots\cdots, y_{n_2}$ be two independent random samples of sizes n_1 and n_2 drawn from two populations with pdf $f(x)$ and $f(y)$ respectively.

If we are interested to test the equality of two populations, we set up the statistical hypothesis as

H_0 : The two populations are identical

i.e., $H_0 : f(x) = f(y)$

Vs

H_1 : The two populations are not identical

i.e., $H_1 : f(x) \neq f(y)$.

The test procedure is based on the pattern of the $x's$ and $y's$ in the combined ordered sample. Let T_1 denote the sum of ranks of $y's$ (in the case when y preceeds x) in the combined ordered sample. The test statistic U defined in terms of T_1 is as follows:

$$U = \left(n_1 n_2 + \frac{n_2(n_2+1)}{2} \right) - T_1 \cdot$$

Similarly, if the ranks of $x's$ are counted (the case when x precedes y), the value of U^1 can be obtained by the formula,

$$U^1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_2$$

Where T_2 is the sum of ranks of $x's$ in the combined sequence of $x's$ and $y's$. There exists relation between U and U^1 which is given as

$$U^1 = n_1 \, n_2 - U \cdot$$

This relation helps to obtain U^1 when U is known and vice - versa without redoing the whole calculations.

Particularly, for n_1 and $n_2 > 8$, standard normal test criteria is fairly good. It has been established that under H_0 U is asymptotically normally distributed with mean $E(U) = \frac{n_1 n_2}{2}$ and variance $V(U) = \frac{n_1 n_2 \left(n_1 + n_2 + 1\right)}{12}$.

The normal test criteria is to compute

$$Z = \frac{U - E(U)}{S.E.(U)} \sim N(0,1)$$

If the calculated value of $|Z| <$ tabulated value of Z at $\alpha\%$ l.o.s., we accept H_0 and conclude that the two populations are identical, otherwise reject H_0.

Problem 1: In order to compare the breaking strength of Nylon fiber produced by two different manufactures, 10 measurments from one producer, 13 measurments from another producer were taken and the results were as follows.

Fiber X :	1.7 1.9 1.8 1.1 0.7 0.9 2.1 1.6 1.7 1.3
Fiber Y :	2.1 2.7 1.6 1.8 1.7 1.8 1.6 2.2 2.4 1.3 1.9 1.8 2.0

Do the data give the evidence that there is a significant difference in the breaking strengths of two Nylon fibers.

Solution: The hypothesis to be tested for the above is

H_0 : There is no significant difference in the breaking strengths of two Nylon fibers.

Vs

H_1 : There is a significant difference in the breaking strengths of two Nylon fibers.

Let us combine both the samples and arrange in the ascending order

Rank :	1	2	3	4.5	4.5	7	7	7	10	10	10	13.5
	X	X	X	X	Y	X	Y	Y	X	X	Y	X
	0.7	0.9	1.1	1.3	1.3	1.6	1.6	1.6	1.7	1.7	1.7	1.8

Rank :	13.5	13.5	13.5	16.5	16.5	18	19.5	19.5	21	22	23
	Y	Y	Y	X	Y	Y	X	Y	Y	Y	Y
	1.8	1.8	1.8	1.9	1.9	2.0	2.1	2.1	2.2	2.4	2.7

T_1 = Sum of ranks of Y's = 189

Since, $n_1,\ n_2 > 8$ the test criteria is to compute the standard normal variate Z and compare with the critical value at 5% l.o.s.

$$E(U) = \frac{n_1 n_2}{2} = \frac{10 \times 13}{2} = 65$$

$$V(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{10 \times 13(10 + 13 + 1)}{12} = 260$$

$$U = \left[n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} \right] - T_1$$

$$= \left[10 \times 13 + \frac{13(13+1)}{2} \right] - 189$$

$$U = 32$$

$$Z = \frac{U - E(U)}{S.E.(U)} = \frac{32 - 65}{\sqrt{260}} = \frac{-33}{16.1245} = -2.0465$$

$$|Z| = 2.0465$$

Table value of Z at 5% l.o.s. for a two - tailed test is 1.96.

Since the calculated $|Z|$ is greater than the tabulated Z at 5% l.o.s. we reject H_0 and conclude that there is a significnat difference in the breaking strengths of the two Nylon fibers.

Problem 2: Given two samples for the study of a particular character, test whether the two samples have been drawn from the same population using Mann - Whitney $'U'$ test.

Sample I:	25 30 28 34 24 25 13 32 34 30 31 35
Sample II:	44 34 22 08 47 31 40 30 32 35 18 21 35 35 29 22

Solution: The hypothesis to be tested is

H_0 : The samples have been drawn from the same population.

Vs

H_1 : The samples are not drawn from the same population.

Combine both the samples arrange them in ascending order and rank them.

Rank:	1	2	3	4	5.5	5.5	7	8.5	8.5	10	11	13	13	13	15.5	15.5
	Y	X	Y	Y	Y	Y	X	X	X	X	Y	X	X	Y	X	Y
	08	13	18	21	22	22	24	25	25	28	29	30	30	30	31	31

Rank:	17.5	17.5	20	20	20	23.5	23.5	23.5	23.5	26	27	28
	X	Y	X	X	Y	X	Y	Y	Y	Y	Y	Y
	32	32	34	34	34	35	35	35	35	40	44	47

T_1 = Sum of ranks of Y = 247.5

$$U = \left[n_1 n_2 + \frac{n_2(n_2+1)}{2}\right] - T_1$$

$$= \left[12 \times 16 + \frac{16(16+1)}{2}\right] - 247.5$$

= 80.5

As $n_1, n_2 \geq 8$ the test criteria is to apply normal test.

$$Z = \frac{U - E(U)}{V(U)} \sim N(0,1) \text{ at } \alpha \text{ \% l.o.s.}$$

$$E(U) = \frac{n_1 n_2}{2} = \frac{12 \times 16}{2} = 96$$

$$V(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{12 \times 16 (12 + 16 + 1)}{12} = 464$$

$$\text{S.E.}(U) = 21.54$$

$$Z = \frac{80.5 - 96}{21.54} = -0.719$$

$$|Z| = 0.719$$

Tabulated value of Z at 5% l.o.s. for two tailed test is 1.96.

Since, the calculated $|Z| <$ tabulated value at 5% l.o.s., we accept H_0 and conclude that the samples have been drawn from the same population.

16.4 Median Test:

Median test is attributed to Westenberg (1948) and Mood (1950). The sign test for two sample problem has been applicable only when observations are paired, which is not always possible. So, if we have two samples of different sizes, median test is a test ofequality of location parameters of two populations under consideration (or) if two independent samples have been drawn from the population, with the same median.

Let $x_1, x_2, ..., x_{n_1}$ and $y_1, y_2, \cdots\cdots\cdots\cdots\cdots, y_{n_2}$ be two random samples of sizes n_1 and n_2 drawn from two populations with the functions $F(x)$ and $F(y)$ respectively. If we are interested to test the equality of two populations with the same median (or) to test the equality of location parameter (here median) of two populations under consideration, we set up the hypothesis as

$$H_0 \quad : \quad F_X(x) = F_Y(x) \ \forall \ x$$

Vs

$$H_1 \quad : \quad F_X(x) = F_Y(x - \delta) \ \forall \ x \ \text{ and } \ \delta \neq 0$$

where δ is the shift in the location parameter which is the median in the test.

The test procedure for median test is to combine the samples and arrange them in order. Then find the median $'\theta'$ for the ordered sample. Count number of $x's$ and $y's$ on the left to $'\theta'$. Let $'U'$ represents number of $x's$ left to θ and V represents number of $y's$ left to θ. The test is based on u, the number of observations which are less than θ in the combined sample is called Median Test. Under H_0, the probability of $U = u$ is

$$f_U(u) = \frac{\left({}^{n_1}C_u\right)\left({}^{n_2}C_v\right)}{\left({}^{n_1+n_2}C_t\right)} \text{ for } u = 0,1,..........,n_1, \ \ t = \frac{n}{2} \text{ and } n = n_1 + n_2 .$$

The test criteria is to calculate $f_u(u)$ and compare it with α, predetermined level of significance.

For two tailed test if $f_u(u) \le \alpha/2$, reject H_0 and conclude that there is a significance difference in the location parameters of two populations. Otherwise accept H_0.

For one - tailed test if $f_u(u) \le \alpha$, reject H_0. Otherwise accept H_0.

For $n \ge 10$, the test is a standard test. The variab;e $u \sim N\left[E(u), V(u)\right]$, where

$$E(U) = \frac{n_1 t}{n} \quad \text{and} \quad V(U) = \frac{n_1 n_2 (n-t)}{n^2 (n-1)}$$

Hence, the normal test to test H_0 is

$$Z = \frac{U - \frac{n_1 t}{n}}{\sqrt{\frac{n_1 n_2 (n-t)}{n^2 (n-1)}}} \sim Z(0,1) \text{ at } \alpha\% \text{ l.o.s.}$$

If the calculated $|Z|$ is less than the tabulated Z at $\alpha\%$ l.o.s., we accept H_0 and conclude that there is no significance difference in the location parameters of two populations or the two samples have drawn from the two populations with the same median. Otherwise, reject H_0.

Problem 1: The pulse rates of 6 persons without any medication and of 7 persons after 3 days of medication were as follows:

Pulse Rate (X) : **(Without Mediaction)**	120	104	72	182	88	96	
Pulse Rater After: **3 days medication (y)**	122	108	105	130	140	136	84

Test whether the distribution of pulse rates of persons before and after medication is same at $\alpha = 0.05$ using Median test.

Solution: The hypothesis to bet ested for the above is

H_0 : There is no significance difference in the pulse rates of persons before and after medication.

i.e., $H_0 \ : F_X(x) = F_Y(x)$

Vs

H_1 : There is a significance difference in the pulse rates of persons before and after medication.

i.e., $H_1 \ : F_X(x) = F_Y(X - \delta)$

Let us first combine the two samples and arrange them in ascending order.

X Y X X X Y X X Y Y Y Y X

72, 84, 88, 96, 104, 105, 108, 120, 122, 130, 136, 140, 182.

The median for the above data is 108.

$u =$ number of X's left to median = 4

$v =$ number of Y's left to median = 2

$$n_1 = 6 \ \ ; n_2 = 7 \ ; \ n = n_1 + n_2 = 13 \ \ ; \ \ t = \frac{n}{2} = \frac{13}{2} = 6.5 \simeq 7$$

The probability

$$f_u(U) = \frac{\binom{n_1}{u}\binom{n_2}{v}}{\binom{n_1+n_2}{t}} \text{ for } u = 0,1,2,\ldots\ldots\ldots,n \ \ ; \ \ t = \frac{n}{2}$$

$$= \frac{\binom{6}{4}\binom{7}{2}}{\binom{13}{7}} = 0.1835$$

Since calculated probability is greater than the level of significance i.e., 0.184 > 0.05, we accept H_0 and conclude that the distribution of pulse rates before and after medication are same.

Problem 2: The following are the 10 plots each, under two treatments. Test the equality of median response under the two treatments by Median test.

Treatment1 (X) :	46,	45,	32,	42,	39,	48,	49,	30,	51,	34.
Treatment2 (Y) :	44,	40,	59,	47,	55,	50,	47,	71,	43,	55.

Solution: The hypothesis to be tested forthe above is

H_0 : There is no significant difference in the median response of two treatments.

i.e., $H_0 : F_X(x) = F_Y(x)$

Vs

H_1 : There is a significant difference in the median response of two treatments.

i.e., $H_1 : F_X(x) = F_Y(X - \delta)$

Let us combine both the samples and arrange in the ascending order.

X	X	X	X	Y	X	Y	Y	X	X	Y	Y	X	X	Y	X	Y	Y	Y	Y
30	32	34	39	40	42	43	44	45	46	47	47	48	49	50	51	55	55	59	71

Here the size of the combined sample $n = n_1 + n_2 = 10 + 10 = 20$

The median for the given data is 46.5

u = number of X's left to median = 7

v = number of Y's left to median = 3

$$t = \frac{n}{2} = \frac{20}{2} = 10$$

Since, $n \geq 10$, the test is a standard test.

$$Z = \frac{U - E(U)}{\sqrt{V(U)}} \sim N(0, 1)$$

where

$$E(U) = \frac{n_1 t}{n} \quad \& \quad V(U) = \frac{n_1 n_2 (n - t)}{n^2 (n - 1)}$$

$$E(U) = \frac{n_1 t}{n} = \frac{10 \times 10}{20} = \frac{100}{20} = 5$$

$$V(U) = \frac{10 \times 10\,(20-10)}{100(20-1)} = \frac{1000}{1900} = 0.5263$$

$$S.E.(U) = 0.7254$$

$$Z = \frac{7-5}{0.7254} = \frac{2}{0.7254} = 2.7570$$

Tabulated value of Z at 5% l.o.s. for two tailed test is 1.96

Since, the calculated Z is greater than the tabulated Z at 5% l.o.s., we reject H_0 and conclude that there is a significant difference in the median response of two treatments.

16.5 Exercises:

1. The following are the scores obtained for 6 clerks in a parliament and 7 clerks in secretariat in an eligibility test for their promotion.

Scores of clerks in parliament:	40	35	52	60	46	55	
Scores of clerks in secretariat:	47	56	42	57	50	67	62

Test whether the performance of clerks in two offices are equal in the test for their promotion.

2. The following are the rates of flow of a certain gas through two soil samples collected from two different places.

Sample X :	23	27	19	24	22	30		
Sample Y :	21	29	34	32	26	28	36	26

Test whether the populations of soil types are the same with respect to the rates of flow through the soils, using suitable N - P tests.

3. The following are the scores of certain randomly selected students at mid term and final examinations.

Mid Score (X) :	55	57	72	90	57	74	
Final Score (Y) :	80	76	63	58	56	37	75

Test whether the distribution of scores of two occasions are same with the help of Mann - Whitney U test.

4. The school children taking coaching in two private schools secured the following scores out of 100.

School 1 :	33	38	39	48	58	70	61	41	45	49
School 2 :	32	15	87	32	22	63	56	57	44	

Test the hypothesis that the students studying in private schools have identical distribution of marks applying Median test at 1% l.o.s.

5. The quality control laboratories independently collected samples of 25 articles from a number of sales depots and tested them. The number of defectives per sales depot were as follows:

Lab A :	9	3	1	3	0	7	2	11
Lab B :	12	6	6	4	8	5	4	

Test the hypothesis that the two laboratories have samples from the same lot by (i) Median Test, (ii) Mann - Whitney U - Test, (iii) Wald - Wolfowitz Run Test.

16.6 Summary:

This lesson explains the non - parametric tests for two independent samples, particularly, median test, Wald - Wolfowitz Run Test, Wilcoxon - Mann - Whitney U Test. A good number of problems are solved to explain the above test procedures and exercises are given to students to solve on their own.

16.7 Technical Terms:

Wald - Wolfowitz Run Test

Wilcoxon - Mann - Whitney U Test

Median Test

Lesson Writer

P. Nagamani

Practical - 1

FITTING OF STRAIGHT LINE AND PARABOLA

Problem 1:

Fit a straight line using the method of least squares for the following data.

X	1	2	3	4	5	6	7	8	9	10
Y	52.5	58.7	65	70.2	75.4	81.1	87.2	95.5	102.2	108.4

Aim:

To fit a straight line to the given data using method of least squares.

Procedure:

Let $y = a + bx \cdots\cdots (1)$ be the straight line to be fitted to the given data. By the method of least squares the normal equations for estimating 'a' and 'b' are as follows:

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

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

$\sum x, \sum y, \sum x^2, \sum xy$ values are obtained from the given data and 'n' is the number of paired values. After substituting these values in the above normal equations and solving them, we will get the values of a and b. With these estimated values of a and b equation (1) is the best fitted straight line.

Calculations:

To get $\sum x, \sum y, \sum x^2, \sum xy$ the following table is prepared.

	x	y	xy	x^2
	1	52.5	52.5	1
	2	58.7	117.4	4
	3	65.0	195.0	9
	4	70.2	280.8	16
	5	75.4	377.0	25
	6	81.1	486.6	36
	7	87.2	610.4	49
	8	95.5	764.0	64
	9	102.2	919.8	81
	10	108.4	1084.0	100
Total	$\sum x = 55$	$\sum y = 796.2$	$\sum xy = 4887.5$	$\sum x^2 = 385$

Substituting these values in the above normal equations we get

$$796.2 = 10a + 55b$$

$$4887.5 = 55a + 385b$$

Solving these two equations we get a = 79.62 and b = 3.08

$\therefore$ The best fitted straight line to the given data is

$$y = 79.62 + 3.08x$$

Inference:

The best fitted straight line to the given data by using the method of least squares is $y = 79.62 + 3.08x$.

Problem 2:

Fit a second degree parabola to the following data, using x as the independent variable.

X	1	2	3	4	5	6	7	8	9
y	2	6	7	8	10	11	11	10	9

Aim:

To fit a second degree parabola to the given data by using the method of least squares.

Procedure:

Let $y = a + bx + cx^2 \cdots\cdots\cdots\cdots(1)$ be the second degree parabola to be fitted to the given data. By the method of least squares the normal equations for estimating a, b and c are as follows:

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

$n, \sum x, \sum y, \sum xy, \sum x^2, \sum x^2y, \sum x^3, \sum x^4$ values can be obtained from the given data. Solving the above equations we will get the required values of a, b and c. With these values equation (1) is the best fitted second degree parabola.

Calculations:

To get $\sum x, \sum x^2, \sum x^3, \sum x^4, \sum y, \sum xy, \sum x^2y$ the following table is prepared.

	x	y	XY	x^2y	x^2	x^3	x^4
	1	2	2	2	1	1	1
	2	6	12	24	4	8	16
	3	7	21	63	9	27	81
	4	8	32	128	16	64	256
	5	10	50	250	25	125	625
	6	11	66	396	36	216	1296
	7	11	77	539	49	343	2406
	8	10	80	640	64	512	4096
	9	9	81	729	81	729	6561
Total	$\sum x = 45$	$\sum y = 74$	$\sum xy = 421$	$\sum x^2y =$ 2771	$\sum x^2 =$ 285	$\sum x^3 =$ 2025	$\sum x^4 =$ 15333

Substituting these values in the normal equations, we have

$$74 = 9a + 45b + 285c$$

$$421 = 45a + 285b + 2025c$$

$$2771 = 285a + 2025b + 1533c$$

Solving these equations we get

$$a = -1, \; b = 3.55, \; c = -0.27$$

The best fitted parabola is $y = -1 + 3.55x - 0.27x^2$

Inference:

The best fitted second degree parabola to the given data by the method of least squares is

$$y = -1 + 3.55x - 0.27x^2$$

Lesson Writer

A. Mohan Rao

Practical - 2

FITTING OF POWER CURVE AND EXPONENTIAL CURVE

Problem 1:

Fit a power curve of the type $y = ax^b$ for the following data:

X	11	12	13	14	15	16	17	18	19	20
Y	16.2	18.3	25.4	33.4	75.2	78.7	82.2	88.4	95.4	100.7

Aim:

To fit a power curve to the given data by the method of least squares.

Procedure:

$y = ax^b \cdots\cdots\cdots (1)$ be the power curve be fitted to the given data. Taking logarithms on both sides.

$$\log\, y = \log\, a + b \log x$$

$$U = A + bV \cdots\cdots\cdots\cdots (2)$$

where $U = \log\, y$; $A = \log\, a$, $V = \log\, x$.

Now equation (2) is the linear equation in U and V, by the method of least squares the normal equations for estimating A and b are

$$\sum U = nA + b\sum V$$

$$\sum UV = A\sum V + b\sum V^2$$

$\sum U, \sum V, \sum UV, \sum V^2$ values are obtained from the data and solving the above equations we get A and b. a = Antilog A. Substituting these values of a & b in equation (1) we get required equation.

Calculations:

x	y	$U = \log\ y$	$V = \log\ x$	UV	V^2
11	16.2	1.2095	1.0414	1.2595	1.0845
12	18.3	1.2625	1.0792	1.3624	1.1646
13	25.4	1.4048	1.1139	1.5648	1.2407
14	33.4	1.5237	1.1461	1.7463	1.3135
15	75.2	1.8762	1.1761	2.2065	1.3832
16	78.7	1.8960	1.2041	2.2829	1.4498
17	82.2	1.9149	1.2304	2.3560	1.5138
18	88.4	1.9465	1.2553	2.4434	1.5757
19	95.4	1.9795	1.2788	2.5313	1.6353
20	100.7	2.0027	1.3010	2.6057	1.6926
Total		$\sum U = 17.0165$	$\sum V = 11.8263$	$\sum UV = 20.3588$	$\sum V^2 = 14.0537$

Substituting these values in the normal equations, we get

$$17.0165 = 10A + 11.8263b$$

$$20.3588 = 11.8263A + 14.0537b$$

Solving the above equations, we get $b = 3.4674;\ A = -2.399$.

$$a = \text{Anti}\log\ A = \text{Anti}\log(-2.399) = 0.002506$$

$\therefore$ Best fitted power curve is $y = (0.0025)\ x^{3.4674}$

Inference:

The best fitted power curve to the given data by the method of least squares is

$$y = (0.0025)x^{3.4674}$$

Problem 2:

Fit an exponential curve of the type $y = ab^x$ for the following data by the method of least squares.

X	2	3	4	5	6
Y	144	172.8	207.4	248.8	298.6

Aim:

To fit an exponential curve of the type $y = ab^x$ to the given data by the method of least squares.

Procedure:

$y = ab^x \cdots\cdots(1)$ is the exponential curve to be fitted to the given data. Taking logarithms on both sides we get

$$\log y = \log a + x \log b$$

$$\Rightarrow \; U = A + Bx \cdots\cdots(2)$$

where $U = \log y;\ A = \log a;\ B = \log b$.

Equation (2) is a linear equation in U & V. By the method of least squares the normal equations are

$$\sum U = nA + B\sum x$$

$$\sum Ux = A\sum x + B\sum x^2$$

$\sum x,\ \sum U,\ \sum Ux,\ \sum x^2$ values can be obtained from the given data.

Calculations:

$\sum x,\ \sum U,\ \sum Ux,\ \sum x^2$ values can be obtained from the following table:

	x	y	$U = \log y$	Ux	x^2
	2	144.0	2.1584	4.3168	4
	3	172.8	2.2375	6.7125	9
	4	207.4	2.3168	9.2692	16
	5	248.8	2.3959	11.9795	25
	6	298.6	2.4751	14.8506	36
Total	$\sum x = 20$		$\sum U = 11.5837$	$\sum Ux = 47.1266$	$\sum x^2 = 90$

Substituting these values in the normal equations we get

$$11.5837 = 5A + 20B$$

$$47.1266 = 20A + 90B$$

Solving the above equations, we get

A = 1.9995 and B = 0.0792

a = Antilog A = Antilog (1.9995) = 99.88

b = Antilog B = Antilog (0.0792) = 0.12

$\therefore$ Best fitted exponential curve is $y = (99.88)\,(0.12)^x$

Inferenece:

The best fitted exponential curve to the given data by the method of least squares is

$$y = (99.88)\,(0.12)^x$$

Lesson Writer

A. Mohan Rao

Practical - 3

COMPUTATION OF CORRELATION COEFFICIENT, FORMING REGRESSION LINES FOR UNGROUPED DATA AND GROUPED DATA

Problem 1:

Calculate Karl Pearson coefficient of correlation and form the regression lines for the following ungrouped data.

x	45	53	74	82	38	70
y	52	38	73	48	74	90

Aim:

To calculate Karl Pearson coefficient of correlation and form the regression lines for the given data.

Procedure:

For the given bivariate data, Karl Pearson coefficient of correlation

$$r = \frac{\text{Cov}\,(x, y)}{\sigma_x \cdot \sigma_y},$$

where $\text{Cov}(x, y) = \frac{1}{n}\sum xy - \bar{x}\,\bar{y}$

$$\sigma_x = \sqrt{\frac{1}{n}\sum x^2 - \left(\bar{x}\right)^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum y^2 - \left(\bar{y}\right)^2}$$

$$\bar{x} = \frac{\sum x}{n}\,;\ \bar{y} = \frac{\sum y}{n}.$$

Regression Lines:

Regression line of y on x is $y = \bar{y} + r\frac{\sigma_y}{\sigma_x}\left(x - \bar{x}\right)$

Regression line of x on y is $x = \bar{x} + r\frac{\sigma_x}{\sigma_y}\left(y - \bar{y}\right)$

Calculations:

For the above formula we require $\sum x, \sum y, \sum xy, \sum x^2, \sum y^2$ values can be obtained from the following table:

	x	y	x^2	y^2	xy
	45	52	2025	2704	2340
	53	38	2805	1444	2014
	74	73	5476	5329	5402
	82	48	6724	2304	3936
	38	74	1444	5476	2812
	70	90	4900	8100	6300
Total	$\sum x = 362$	$\sum y = 375$	$\sum x^2 = 23378$	$\sum y^2 = 25357$	$\sum xy = 22804$

$$\bar{x} = \frac{\sum x}{n} = \frac{362}{6} = 60.3; \quad \bar{y} = \frac{\sum y}{n} = \frac{375}{6} = 62.5$$

$$\sigma_x = \sqrt{\frac{23378}{6} - (60.3)^2} = 16.13$$

$$\sigma_y = \sqrt{\frac{25357}{6} - (62.5)^2} = 17.88$$

$$\text{Cov}(x, y) = \frac{22804}{6} - (60.3)(62.5) = 31.9$$

Karl Pearson coefficient of correlation $(r) = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} = \frac{31.9}{16.13 \times 17.88}$

$$= 0.11$$

Regression equation of x on y is

$$x = \bar{x} + r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$= 60.3 + 0.11\left(\frac{16.13}{17.88}\right)(y - 62.5)$$

$$\boxed{x = 0.1y + 54.05}$$

Regression equation of y on x is

$$y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$= 62.5 + 0.11\left(\frac{17.88}{16.13}\right)(x - 60.3)$$

$$\boxed{y = 0.12x + 55.25}$$

Inference:

For the given data

Karl Pearson coefficient of correlation (r) = 0.11

Regression equation of x on y is $x = 0.1y + 54.05$

Regression equation of y on x is $y = 0.12x + 55.25$

Problem 2:

Calculate the Karl Pearson coefficient of correlation and form the regression lines for the following grouped data.

y \ x	15 - 25	25 - 35	35 - 45	45 - 55	55 - 65
15 - 30	30	6	3	-	-
30 - 45	18	32	15	12	8
45 - 60	2	28	40	16	9
60 - 75	-	4	9	10	8

Aim:

To calculate Karl Pearson coefficient of correlation and form the regression lines to the given data.

Procedure:

Karl Pearson coefficient of correlation

$$(r) = \frac{N\sum f\,dxdy - \left(\sum fdx\right)\left(\sum fdy\right)}{\sqrt{N\sum fdx^2 - \left(\sum fdx\right)^2}\ \sqrt{N\sum fdy^2 - \left(\sum fdy\right)^2}}$$

$$A \cdot M\left(\bar{x}\right) = A + \frac{\sum fdx}{N} \times h \text{ ; } h = C \cdot I \text{ of } x \text{ series}$$

$$A \cdot M\left(\bar{y}\right) = B + \frac{\sum fdy}{N} \times k \text{ ; } k = C \cdot I \text{ of } y \text{ series}$$

Regression line of y on x is $y = \bar{y} + r\,\frac{\sigma_y}{\sigma_x}\left(x - \bar{x}\right)$

Regression line of x on y is $x = \bar{x} + r\,\frac{\sigma_x}{\sigma_y}\left(y - \bar{y}\right)$

where $\sigma_x = h\sqrt{\frac{\sum fd_x^2}{N} - \left(\frac{\sum fdx}{N}\right)^2}$

$$\sigma_y = k\sqrt{\frac{\sum fd_y^2}{N} - \left(\frac{\sum fdy}{N}\right)^2}$$

Calculations:

To find all the required values in the formula the following table is prepared.

$$dx = \frac{x - A}{h} = \frac{x - 40}{10} \text{ ; } dy = \frac{y - B}{k} = \frac{y - 37.5}{15}$$

x y	**15 - 25**	**25 - 35**	**35 - 45**	**45 - 55**	**55 - 65**	**f**	$dy = \frac{y-B}{k}$	fdy	fd^2y	$fdxdy$
15 - 30	30	6	3	-	-	39	-1	-39	39	66
30 - 45	18	32	15	12	8	85	0	0	0	0
45 - 60	2	28	40	16	9	95	1	95	95	2
60 - 75	-	4	9	10	8	31	2	62	124	44
f	50	70	67	38	25	250		118	258	112
$dx = \frac{x-A}{h}$	-2	-1	0	1	2					
fdx	-100	-70	0	38	50	-82				
fdx^2	200	70	0	38	100	408				
$fdxdy$	56	-30	0	36	50	112				

$N = \sum f = 250;\ \sum fdy = 118; \sum fdx = -82;\ \sum fd^2x = 408;$

$\sum fd_y^2 = 258;\ \sum fdxdy = 112$.

Karl Pearson correlation coefficient $(r) = \dfrac{250 \times 112 - (-82)(118)}{\sqrt{250 \times 408 - (-82)^2}\ \sqrt{250 \times 258 - (118)^2}}$

$= 0.542$

$$\bar{x} = 40 + \frac{(-82)}{250} \times 10 = 36.72$$

$$\bar{y} = 37.5 + \frac{118}{250} \times 15 = 44.58$$

$$\sigma_x = 10\sqrt{\frac{408}{250} - \left(\frac{-82}{250}\right)^2} = 12.35$$

$$\sigma_y = 15\sqrt{\frac{258}{250} - \left(\frac{118}{250}\right)^2} = 13.49$$

Regression line of y on x is

$$y = 44.58 + 0.542\left(\frac{13.49}{12.35}\right)(x - 36.72)$$

$$y = 22.841 + 0.592x$$

Regression line of x on y is

$$x = 36.72 + 0.542\left(\frac{12.35}{13.49}\right)(y - 44.58)$$

$$x = 14.6 + 0.496\ y$$

Inference:

Karl Pearson coefficient of correlation for the given data $(r) = 0.542$

Regression line of y on x is $y = 0.592x + 22.841$

Regression line of x on y is $x = 0.496y + 14.6$

Lesson Writer

A. Mohan Rao

Practical - 4

COMPUTATION OF MULTIPLE AND PARTIAL CORRELATION COEFFICIENTS

Problem 1:

The simple correlation coefficients between temparature (x_1) corn yield (x_2) and rain fall (x_3) are $r_{12} = 0.59;\ r_{13} = 0.46$ and $r_{23} = 0.77$. Calculate the partial correlation coefficients $r_{12.3},\ r_{23.1}$ and $r_{31.2}$. Also calculate multiple correlation coefficient $R_{1.23}$.

Aim:

To calculate partial correlation coefficients. $r_{12.3}, r_{23.1}$ and $r_{31.2}$ and also to calculate multiple correlation coefficient $R_{1.23}$.

Procedure:

Given the simple correlation coefficients r_{12}, r_{13} and r_{23}, the partial correlation coefficients and multiple correlation coefficient are calculated as follows:

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

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{\left(1-r_{21}^2\right)\left(1-r_{31}^2\right)}}$$

$$r_{31.2} = \frac{r_{31} - r_{32}r_{12}}{\sqrt{\left(1-r_{32}^2\right)\left(1-r_{12}^2\right)}}$$

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

Calculations:

We are given $r_{12} = 0.59;\ r_{13} = 0.46,\ r_{23} = 0.77$

Partial correlation coefficients:

$$r_{12.3}=\frac{0.59-0.46\times0.77}{\sqrt{\left[1-(0.46)^2\right]\left[1-(0.77)^2\right]}}$$

$$=0.42$$

$$r_{23.1}=\frac{0.77-0.59\times0.46}{\sqrt{\left[1-(0.59)^2\right]\left[1-(0.46)^2\right]}}=0.69$$

$$r_{31.2}=\frac{0.46-0.59\times0.77}{\sqrt{\left[1-(0.59)^2\right]\left[1-(0.77)^2\right]}}=0.01$$

Multiple correlation coefficient:

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

$$=\sqrt{\frac{0.59^2+0.46^2-2\times0.59\times0.46\times0.77}{1-0.77^2}}$$

$$=0.59$$

Inference:

For the given information partial correlation coefficients are $r_{12.3}=0.42,\ r_{23.1}=0.69,$ $r_{31.2}=0.01$, and the multiple correlation coefficient $R_{1.23}=0.59$.

Lesson Writer

A. Mohan Rao

Practical - 5

COMPUTATION OF CORRELATION RATIO

Problem 1:

Calculate correlation ratio of Y on X $\left(\eta_{yx}\right)$ for the following data:

x y	**10**	**15**	**20**	**25**
7	3	2	-	-
9	-	1	4	6
11	-	3	4	2
13	2	1	5	-
15	-	6	-	1

Aim:

To calculate correlation ratio of Y on X $\left(\eta_{yx}\right)$ for the given data.

Procedure:

Correlation ratio of Y on X $\left(\eta_{yx}\right)$ can be calculated as follows:

$$\eta_{yx}^{2} = \frac{\left[\sum_{i}\left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right]}{\left[\sum n_j y_j^2 - \frac{T^2}{N}\right]}$$

where $T_i = \sum_{j} f_{ij} y_{ij}$

$$\eta_i = \sum_{j} f_{ij},\ n_j = \sum_{i} f_{ij},\ T = \sum_{i} T_i \ .$$

Calculations:

The above values are obtained from the following data.

y \ x	**10**	**15**	**20**	**25**	n_j	$n_j y_j$	$n_j y_j^2$
7	3	2	-	-	5	35	245
9	-	1	4	6	11	99	891
11	-	3	4	2	9	99	1089
13	2	1	5	-	8	104	1352
15	-	6	-	1	7	105	1575
n_i	5	13	13	9	40	442	5152
T_i	47	159	145	91	442	4923.9111	
$\frac{T_i^2}{n_i}$	441.8	1944.6928	1617.3077	920.1111	4923.9111		

$$T = 442;\ \ N = 40;\quad \sum n_j y_j^2 = 5152\ ;\ \sum_i \frac{T_i^2}{n_i} = 4923.9111$$

$$\eta_{yx}^2 = \frac{\left[\sum\left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right]}{\left[\sum n_j y_j^2 - \frac{T^2}{N}\right]} = \frac{4923.9111 - \frac{(442)^2}{40}}{5152 - \frac{(442)^2}{40}} = 0.1486$$

$$\eta_{yx} = \sqrt{0.1486} = 0.3855$$

Correlation ratio of Y on X is η_{yx} is 0.3855

Inference:

Correlation ratio of Y on X $\left(\eta_{yx}\right)$ to the given data is 0.3855.

Lesson Writer

A. Mohan Rao

Practical - 6

TEST FOR MEANS AND S.D.'s

Problem 1:

Given below are the gain in weights of calves on two different diets A and B. Test whether there is any significant difference in the mean weight of calves of two diets.

A:	30	28	27	22	50	71	9	68	10	83	78
B :	20	5	35	43	13	13	60	43	78	88	60
A:	41	67	72	67	22	94	19	43	28	72	41
B :	23	31	94	26	59	70	52	5	20	19	3
A:	6	16	77	93	15	54	89	47	24		
B :	82	59	9	44	30	90	80	29			

Aim:

To test whether there is any significant difference in the mean weights of calves fed on two diets A and B.

Procedure:

Let $\bar{x}_1$ be the sample mean of n_1 observations drawn at random from a population of diet A with mean μ_1 and variance σ_1^2.

Let $\bar{x}_2$ be the sample mean of sample size n_2 observations drawn at random from a population of diet B with mean μ_2 and variance σ_2^2.

Suppose we want to test whether the two population means μ_1 and $\mu 2$ are equal or whether the two sample means $\bar{x}_1$ and $\bar{x}_2$ are significantly different, we set up the statistical hypothesis as follows:

H_0 : The two population means are equal.

i.e., $H_0 : \mu_1 = \mu_2$

Vs

H_1 : The two population means are not equal.

i.e., $H_1 : \mu_1 \neq \mu_2$

The required test statistic to test above hypothesis is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1) \text{ under } H_0 .$$

where $s_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(x_i - \bar{x}_1\right)^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_1^2 - \left(\bar{x}_1\right)^2$

$$s_2^2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \left(x_j - \bar{x}_2\right)^2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_2^2 - \left(\bar{x}_2\right)^2$$

If the calculated value of $|Z|$ is less than the table value of Z at α % l.o.s. we accept H_0 and conclude that there is no significant difference between the two sample means. Otherwise reject H_0 .

Calculations:

X_1	X_2	X_1^2	X_2^2
30	20	900	400
28	5	784	25
27	35	729	1225
22	43	484	1849
50	13	2500	169
71	13	5041	169
9	60	81	3600
68	43	4624	1849
10	78	100	6084
83	88	6889	7744
78	60	6084	3600
41	23	1681	529
67	31	4489	961
72	94	5184	8836
67	26	4489	676
22	59	484	3481
94	70	8836	4900

19	52	361	2704
43	5	1849	25
28	20	784	400
72	19	5184	361
41	30	1681	900
6	82	36	6724
16	59	256	3481
77	9	5929	81
93	44	8649	1936
15	30	225	900
54	90	2916	8100
89	80	7921	6400
47	29	2209	841
24		576	
1463	1310	91955	82449

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{1463}{31} = 47.1935$$

$$\bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{1310}{30} = 43.666$$

$$s_1^2 = \frac{1}{n_1}\sum x_1^2 - \left(\bar{x}_1\right)^2$$

$$s_1^2 = \frac{91955}{31} - (47.1935)^2$$

$$s_1^2 = 739.0639$$

$$s_2^2 = \frac{1}{n_2}\sum x_2^2 - \left(\bar{x}_2\right)^2$$

$$s_2^2 = \frac{82449}{30} - (43.0666)^2$$

$$s_2^2 = 841.6$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

$$Z = \frac{47.1935 - 43.666}{\sqrt{\frac{739.06}{31} + \frac{841.6}{30}}}$$

$$Z = \frac{3.5269}{7.2034}$$

$$Z = 0.4896$$

Tabulated value of Z for two tailed test at 5% l.o.s., is 1.96 (see Table 13 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

Since calculated value of Z is less than the tabulated value of Z at 5% l.o.s., we accept H_0 and infer that there is no significant difference between the two population means.

Problem 2:

The following two independent samples are drawn from two populations. Test whethere the population standard deviations are equal or not.

Sample I :	64	43	42	60	85	46	7	80	93	25	68
	75	54	68	88	31	43	18	4	58	76	83
	83	25	35	43	69	89	10	53	23	49	14
	26	10									
Sample II :	16	59	27	26	56	88	92	21	25	68	14
	98	55	35	43	80	46	12	72	67	5	64
	35	58	9	27	41	13	43	17	37	64	55
	65	9	6	8	44						

Aim:

To test whether the two sample standard deviations are significant or not.

Procedure:

Let s_1 and s_2 be the standard deviations of two independent samples taken from two populations with their respective standard deviations σ_1 and σ_2. If we are interested to test whether the two population standard deviations are equal, we setup the following hypothesis as

H_0 : There is no significant difference between the two sample s.d's

i.e., $H_0 : \sigma_1 = \sigma_2$

Vs

H_1 : There is a significant difference between the two sample s.d.'s

i.e., $H_1 : \sigma_1 \neq \sigma_2$.

The test statistic to test the hypothesis is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} \sim N(0,1) \text{ under } H_0 .$$

If the calculated value of Z less than the tabulated value of Z at α % l.o.s., we accept H_0 and conclude that there is no significant difference between the two sample s.d.'s. Otherwise reject H_0.

Calculations:

X_1	X_2	X_1^2	X_2^2
64	16	4096	256
43	59	1849	3481
42	27	1764	729
60	26	3600	676
85	56	7225	3136
46	88	2116	7744
7	92	49	8464
80	21	6400	441
93	25	8649	625
25	68	625	4624
68	14	4624	196

75	98	5625	9604
54	55	2916	3025
68	35	4624	1225
88	43	7744	1849
31	80	961	6400
43	46	1849	2116
18	12	324	144
4	72	16	5184
58	67	3364	4489
76	5	5776	25
83	64	6889	4096
83	35	6889	1225
25	58	625	3364
35	9	1225	81
43	27	1849	729
69	41	4761	1681
89	93	7921	8649
10	43	100	1849
53	17	2809	289
23	37	529	1369
49	64	2401	4096
14	55	196	3025
26	65	676	4225
10	9	100	81
	6		36
	8		64
	44		1936
1689	1680	111166	101228

$$\sum x_1 = 1689; \ \sum x_2 = 1680$$

$$\sum x_1^2 = 111166 \ ; \ \sum x_2^2 = 101228$$

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{1689}{35} = 48.2571$$

$$\bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{1680}{38} = 44.2105$$

$$s_1^2 = \frac{1}{n_1}\Sigma x_1^2 - \left(\bar{x}_1\right)^2$$

$$s_1^2 = \frac{111166}{35} - (48.2571)^2$$

$$s_1^2 = 847.4236 \Rightarrow s_1 = 29.1105$$

$$s_2^2 = \frac{1}{n_2}\Sigma x_2^2 - \left(\bar{x}_2\right)^2$$

$$s_2^2 = \frac{101228}{38} - (44.2105)^2$$

$$s_2^2 = 2663.8947 - 1954.5683$$

$$s_2^2 = 709.3264$$

$$s_2 = 26.6331$$

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} \sim N(0,1)$$

$$Z = \frac{29.1105 - 26.6331}{\sqrt{\frac{847.4236}{2 \times 35} + \frac{709.3264}{2 \times 38}}} = \frac{2.4774}{4.6302} = 0.5350$$

Tabulated value of Z at 5% l.o.s. is 1.96. (See table 13 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

Since the calculated value of Z is less than the tabulated value of at 5% level of significance, we accept H_0 and infer that there is no significant difference among the two sample s.d.'s.

Lesson Writer

P. NagaMani

Practical - 7

TEST FOR PROPORTIONS

Problem 1:

In a random sample of 400 persons from a large population, 120 are females. Can it be said that males and females are in the ratio 5 : 3 in the population. Use 1% level of significance.

Aim:

To test whether the males and females are in the ratio of 5 : 3 in the population.

Procedure:

Let 'x' be number of individuals posessing the attribute with a sample of 'n' individuals drawn from a population with P, the population proportion.

Then $X \sim B(n, P)$.

Let $p = \frac{x}{n}$ be the sample proportion of individuals posessing the attribute, then

$$E(p) = \frac{1}{n} E(x) = \frac{nP}{n} = P$$

$$V(p) = \frac{1}{n^2} V(x) = \frac{1}{n^2} nP(1-P) = \frac{P(1-P)}{n}$$

For large samples $X \sim N\left[np, np(1-p)\right]$ and $p \sim N\left[P, \frac{P(1-P)}{n}\right]$

Suppose we want to test the population proportion $P = P_0$, we setup statistical hypothesis as

H_0 : There is no significant difference between the sample proportion and the population proportion.

i.e., $H_0 : P = P_0$

Vs

H_1 : There is a significant difference between the sample proportion and the population proportion.

i.e., $H_1 : P \neq P_0$

The required test statistic to test the above hypothesis is

$$Z = \frac{p - E(p)}{SE(p)} \sim N(0,1) \text{ under } H_0 .$$

$$Z = \frac{p - P}{\sqrt{\frac{P(1-P)}{n}}} \sim N(0,1) \text{ under } H_0 .$$

If the calculated value of $|Z|$ is less than the tabulated value of Z at α % l.o.s., we accept H_0 and we conclude that there is no significant difference between the sample proportion and the population proportion otherwise reject H_0 .

Calculations:

We are given

$n = 400$ and x = number of females in the sample = 120.

p = sample proportion of females

$$p = \frac{120}{400} = 0.30$$

Population proportion of females $P = \frac{3}{8} = 0.375$

The required test statistic to test H_0 is

$$Z = \frac{p - P}{\sqrt{\frac{P(1-P)}{n}}} \sim N(0,1)$$

$$Z = \frac{0.30 - 0.375}{\sqrt{\frac{0.375(1-0.375)}{400}}}$$

$$Z = \frac{-0.075}{0.024} = -3.125$$

$$|Z| = 3.125$$

Tabulated value of Z for two tail test at 1% l.o.s. is 2.58. (see Table 13 of Prof. T.V. Avadhani's

Statistical tables).

Inference :

Since the calculated $|Z|$ is greater than the tabulated Z at 1% l.o.s., we reject H_0 and conclude that males and females in the population are not in the ratio of 5 : 3.

Problem 2:

The subject under investigation is the measure of dependence of Tamil on words of Sanskrit origin. One news paper article reporting the proceedings of the constituent assembly contained 2025 words of which 729 words were declared by literary critic to be of sanskrit origin. A second article by the same author describing automic research contained 1600 words of which 640 words were declared by the same critic to be of Sanskrit origin. Examine whether there is any significant difference in the dependence of this writer on words of Sanskrit origin in writing the two articles.

Aim:

To examine if there is any significant difference in the dependence of a writer on words of Sanskrit origin in writing the two articles.

Procedure:

Let x_1 be the number of individuals posessing an attribute in a sample of $'n_1'$ individuals from a population with proportion P_1.

Let x_2 be the number of individuals posessing the attribute in a sample of n_2 individuals from another population with proportion P_2.

Suppose, if we want to test the two populations are equal, we set up the hypothesis.

H_0 : Two population proportions are equal.

i.e. $H_0 : P_1 = P_2$

Vs

H_1 : Two population proportions are not equal.

i.e., $H_1 : P_1 \neq P_2$

The required test statistic to test the hypothesis is

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1), \text{ where } \hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

If the calculated $|Z|$ is less than the tabulated Z at $\alpha\%$ l.o.s., we accept H_0 and conclude that there is no significant difference between the two sample proportions. Otherwise reject H_0.

Calculations:

We are given

$$n_1 = 2025,\ n_2 = 1600$$

$p_1 =$ sample proportion of the first article.

$$p_1 = \frac{x_1}{n_1} = \frac{\text{Number of Tamil dependence words on Sanskrit in article 1}}{\text{Size of the first article}}$$

$$p_1 = \frac{729}{2025} = 0.36$$

Let x_2 represent number of Tamil dependence words on Sanskrit in article 2.

Therefore, the proportion of the second article is

$$p_2 = \frac{x_2}{n_2} = \frac{640}{1600} = 0.40$$

The required test statistic to test H_0 is

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

where $\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$$\hat{p} = \frac{729 + 640}{2025 + 1600} = 0.38$$

$$1 - \hat{p} = 0.62$$

$$Z = \frac{0.36 - 0.40}{\sqrt{0.38 \times 0.62\left(\frac{1}{2025} + \frac{1}{1600}\right)}} = \frac{-0.04}{0.0162} = -2.469$$

$$|Z| = 2.469$$

Tabulated value of Z at 5% l.o.s. is 1.96. (see Table 13 of Prof. T.V. Avadhani's Statistical Tables).

Calculated value of $|Z|$ is greater than the tabulated value of Z at 5% l.o.s. we reject H_0.

Inference:

There is a significance difference in the dependence of the writer on words of Sanskrit origin in writing the two articles.

Lesson Writer
P. NagaMani

Practical - 8

TEST FOR MEANS AND CORRELATION

Problem 1:

The following is a random sample of 15 measurments of thickness of plastic sheet (in inches) used for making chairs.

5.64, 7.48, 3.99, 6.44, 8.95, 6.48, 7.96, 4.96, 5.98, 2.99, 8.63, 8.24, 7.36, 4.05, 7.77.

Test whether the population mean of thickness of the plastic sheet is 6.45 inches.

Aim:

To test whether the population mean of thickness of the plastic sheet is 6.45 inches or not.

Procedure:

Here the test procedure is of t - test for single mean. Let $x_1, x_2, ..., x_n$ be a random sample of size n, drawn from the population with mean μ and unknown variance. And,

$\bar{x} = \frac{\sum X_i}{n}$ be the sample mean

$s^2 = \frac{1}{n}\sum\left(X_i - \bar{X}\right)^2$ be the sample variance

$S^2 = \frac{1}{n-1}\sum\left(X_i - \bar{X}\right)^2$ be the unbiased sample variance

Then to test the population mean $\mu = \mu_0$ hypothetically we setup the statistical hypothesis as

H_0 : There is no significant difference between population mean and sample mean.

i.e., $H_0 : \mu = \mu_0$

Vs

H_1 : There is a significant difference between population mean and sample mean.

i.e., $H_1 : \mu \neq \mu_0$

The test statistic to test H_0 is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} = \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{(n-1)} \text{ d.f. at } \alpha \text{ \% l.o.s.}$$

If the calculated $|t| <$ tabulated t for $(n-1)$ d.f. at 5% l.o.s., we accept H_0 otherwise reject H_0.

Calculations:

X	$\left(X-\overline{X}\right)$	$\left(X-\overline{X}\right)^2$
5.64	-0.82	0.6724
7.48	1.02	1.0404
3.99	-2.47	6.1009
6.44	-0.02	0.0004
8.95	2.49	6.2001
6.48	0.02	0.0004
7.96	1.5	2.25
4.96	-1.5	2.25
5.98	0.48	0.2304
2.99	-3.47	12.0409
8.63	2.17	4.7089
8.24	1.78	3.1684
7.36	0.9	0.81
4.05	-2.41	5.8081
7.77	1.31	1.7164
$\sum X = 96.92$		$\sum\left(X_i-\overline{X}\right)^2 = 46.9974$

$$\overline{X} = \frac{\sum X}{n} = \frac{96.92}{15} = 6.46$$

$$s = \sqrt{\frac{1}{n}\sum\left(X_i-\overline{X}\right)^2} = \sqrt{\frac{46.9974}{10}} = 1.7700$$

$$t = \frac{\overline{x}-\mu}{s/\sqrt{n-1}} = \frac{6.46-6.45}{1.77/\sqrt{14}} = \frac{0.01}{0.4731} = 0.0211 \sim t_{14} \text{ d.f.}$$

Tabulated value of t for 14 d.f. at 5% l.o.s. is 2.15 (see Table 14 of Prof. T.V. Avadhani's statistical Tables).

Inference:

Since, the calculated value of t is less than the tabulated value of t, we accept H_0 and conclude that the sample data supports the assumption of population mean of thickness is 6.45 inches.

Problem 2:

Two samples of sizes 14 and 21 are drawn from two different populations. The results are as follows:

Sample I : 14, 21, 28, 32, 34, 18, 21, 46, 38, 29, 52, 40, 28, 16.

Sample II : 21, 32, 37, 28, 40, 42, 46, 55, 20, 31, 25, 64, 71, 45, 20, 49, 30, 37, 38, 26, 46.

Test whether the population means are equal or not.

Aim:

To test whether the two population means are equal or not.

Procedure:

The test procedure is t - test for difference of means. Let $X_1, X_2, \ldots\ldots\ldots\ldots\ldots\ldots, X_{n_1}$ be a random sample of size n_1 drawn from a normal population with mean μ_1 and unknown variance σ_1^2. And $\overline{x} = \frac{\sum X_i}{n_1}$ be the sample mean.

$s_1^2 = \frac{1}{n_1}\Sigma\left(X_i - \overline{X}\right)^2$ be the sample variance.

Let $Y_1, Y_2, \ldots\ldots\ldots\ldots\ldots, Y_{n_2}$ be other random sample of size n_2, drawn from another normal population with mean μ_2 and unknown variance σ_2^2. And $\overline{Y} = \frac{\sum Y_j}{n_2}$ be the sample mean.

$s_2^2 = \frac{1}{n_2}\Sigma\left(Y_j - \overline{Y}\right)^2$ be the sample variance.

The statistical hypothesis to test the equality of two population means is

H_0 : There is no significant difference between two sample mean.

i.e., $H_0 : \mu_1 = \mu_2$

Vs

H_1 : There is a significant difference between two sample mean.

i.e., $H_1 : \mu_1 \neq \mu_2$

The test statistic to test H_0 is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \text{ d.f. at 5\% l.o.s.}$$

where $S = \sqrt{\frac{1}{n_1 + n_2 - 2}\left[\Sigma\left(x_i - \bar{x}\right)^2 + \Sigma\left(y_j - \bar{y}\right)^2\right]}$

If the calculated $|t|$ is less than the tabulated value of t for $(n_1 + n_2 - 2)$ d.f. at 5% l.o.s., we accept H_0, otherwise reject H_0.

Calculations:

X	Y	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$
14	21	-15.78	248.06	-17.23	296.87
21	32	-8.78	77.08	-6.23	38.81
28	37	-1.78	3.16	-1.23	1.51
32	28	2.22	4.92	-10.23	104.65
34	40	4.22	17.80	1.77	3.13
18	42	-11.78	138.76	3.77	14.21
21	46	-8.78	77.08	7.77	60.37
46	55	16.22	263.08	16.77	281.23
38	20	8.22	67.56	-18.23	332.33
29	31	-0.78	0.60	-7.23	52.27
52	25	22.22	493.72	-13.23	175.03
40	64	10.22	104.44	25.77	664.09
28	71	-1.78	3.16	32.77	1073.87
16	45	-13.78	189.88	6.77	45.83
	20			-18.23	332.33
	49			10.77	115.99
	30			-8.23	67.73
	37			-1.23	1.51
	38			- 0.23	0.05
	26			-12.23	149.57
	46			7.77	60.37
$\sum x = 417$	$\sum y = 803$		$\Sigma\left(X_i - \bar{X}\right)^2$ $= 1690.24$		$\Sigma\left(Y_j - \bar{Y}\right)^2$ $= 3871.75$

$$\overline{X}=\frac{\sum X_i}{n_1}=\frac{417}{14}=29.78$$

$$\overline{Y}\ \frac{\sum Y_j}{n_2}=\frac{803}{21}=38.23$$

$$S=\sqrt{\frac{1}{n_1+n_2-2}\left[\sum\left(X_i-\overline{X}\right)^2+\sum\left(Y_j-\overline{Y}\right)^2\right]}$$

$$=\sqrt{\frac{1}{14+21-2}[1690.24+3871.75]}$$

$$=\sqrt{\frac{5561.99}{33}}$$

$$S=12.9824$$

$$t=\frac{\overline{x}-\overline{y}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{29.78-38.23}{(12.9824)\sqrt{\frac{1}{14}+\frac{1}{21}}}$$

$$=\frac{-8.45}{(12.9824)(0.3310)}=\frac{-8.45}{4.30}=-1.95$$

$$|t|=1.95$$

Tabulated value of t for 33 d.f. at 5% l.o.s. for two - tailed test is 2.042 (see Table 14 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

Since the calculated value of t is less than the tabulated value of t for 33 d.f. at 5% l.o.s., we accept H_0 and conclude that there is no significant difference between the two sample means.

Problem 3:

The following data gives the distribution of items of production and also the relative defective items among them, according to size - groups. Test whether the percentage of defectives are correlated with the size of the groups.

Size - Group :	15 - 16	16 - 17	17- 18	18 - 19	19 - 20	20 - 21
% of defectives:	75	60	50	50	45	40

Size - Group :	21 - 22	22 - 23	23 - 24	24 - 25
% of devectives:	65	60	35	55

Aim:

To test whether the percentage of defectives are correlated with the size of the groups or not.

Procedure:

The test procedure is t - test for significance of correlation coefficient.

Let $(X_i, Y_i) \forall\ i = 1, 2, ..., n$ be a bi - variate random sample drawn from a bi - variate normal population with correlation coefficient 'ρ'. Let 'r' be the sample correlation coefficient given by

$$r = \frac{\text{Cov}(X, Y)}{\sigma_x \ \sigma_y} = \frac{\frac{1}{n}\sum xy - \bar{x}\ \bar{y}}{\sqrt{\frac{1}{n}\sum x^2 - (\bar{x})^2}\ \sqrt{\frac{1}{n}\sum y^2 - (\bar{y})^2}}$$

If we want to test whether the population correlation coefficient $\rho = 0$ i.e., the variables are uncorrelated, we set up the statistical hypothesis as

H_0 : The variables are uncorrelated.

i.e., $\rho = 0$

Vs

H_1 : The variables are correlated.

i.e., $\rho \neq 0$

The test statistic to test he above hypothesis is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$ d.f. at α % l.o.s.

If the calculated value of $|t|$ < tabulated value of t at $(n-2)$ d.f. at predetermined level of significance α, we accept H_0, otherwise reject H_0.

Calculations:

C.I.	X	Y	X^2	Y^2	XY
15 - 16	15.5	75	240.25	5625	1162.5
16 - 17	16.5	60	272.25	3600	990.0
17 -18	17.5	50	297.56	2500	875.0
18 - 19	18.5	50	342.25	2500	925.0
19 - 20	19.5	45	380.25	2025	877.5
20 - 21	20.5	40	420.25	1600	820.0
21 - 22	21.5	65	462.25	4225	1397.5
22 - 23	22.5	60	506.25	3600	1350.0
23 - 24	23.5	35	552.25	1225	822.5
24 - 25	24.5	55	600.25	3025	1347.5
	$\sum x = 200$	$\sum y = 535$	$\sum x^2 = 4073.81$	$\sum y^2 =$ 29925	$\sum xy =$ 10567.5

$$r = \frac{\frac{1}{n}\sum xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum x^2 - \left(\bar{x}\right)^2}\ \sqrt{\frac{1}{n}\sum y^2 - \left(\bar{y}\right)^2}}$$

$$\bar{x} = \frac{1}{n}\sum x = \frac{200}{10} = 20$$

$$\bar{y} = \frac{1}{n}\sum y = \frac{535}{10} = 53.5$$

$$r = \frac{\frac{1}{10}\times 10567.5 - (20)(53.5)}{\sqrt{\frac{1}{10}\times 4073.81 - (20)^2}\ \sqrt{\frac{1}{10}\times 29925 - (53.5)^2}}$$

$$r = \frac{1056.75 - 1070}{2.7168 \times 11.4127} = \frac{-13.25}{31.0060} = -0.4273$$

The sample correlation coefficient $r = -0.4273$

The required test statistic to test the hypothesis is

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}=\frac{-0.4273\sqrt{10-2}}{\sqrt{1-(0.4273)^2}}=\frac{-1.2086}{0.9041}=-1.3368$$

Tabulated value of t for $n-2=8$ d.f. at 5% l.o.s. for two tailed test is 2.31(see Table 14 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

Since calculated $|t|$ < tabulated t for 8 d.f. at 5% level of significance we accept H_0 and conclude that the percentage of defectives are uncorrelated with the size of the group.

Problem 4:

In a certain experiment two types of foods A and B are fed to the same set of 15 pigs. Below given are the gain in weights of pigs Kgs, can we conclude that food A is better than food B.

Gain in Weight

Diet A(x) :	25,	32,	30,	34,	24,	14,	32,	24,	30,	31,	35,	25,	24,	14,	32
Diet B(y) :	44,	34,	22,	10,	47,	31,	40,	30,	32,	35,	18,	21,	35,	29,	22

Aim:

To test whether the food A is better than food B w.r.t gain in weights of pigs.

Procedure:

Let $X_1, X_2, \cdots\cdots\cdots\cdots\cdots, X_n$ and $Y_1, Y_{2,} \cdots\cdots\cdots\cdots\cdots, Y_n$ be two samples which are not independent, then define $d_i=(X_i-Y_i)\ \forall\ i=1,\ldots\ldots\ldots\ldots,n$ and the hypothesis to be tested is

H_0 : There is no significance difference in food A and food B

i.e., $H_0:\mu_A=\mu_B$

Vs

H_1 : There is a significant difference i.e., food A is better than food B

i.e., $H_1:\mu_A>\mu_B$

The required test statistic to test the hypothesis is

$$t=\frac{\bar{d}}{s/\sqrt{n-1}}\sim t_{(n-1)} \text{ d.f. at } \alpha\% \text{ l.o.s.}$$

where $\bar{d}=\sum d_i/n$ and $s=\sqrt{\frac{1}{n}\sum\left(d_i-\bar{d}\right)^2}$

If the calculated value of $|t|$ is less than the tabulated t with (n - 1) d.f. at $\alpha\%$ l.o.s., we accept H_0 and conclude that there is no significant difference between the two sample means, otherwise reject H_0.

Calculations:

X :	25 31	32 35	30 25	34 24	24 14	14 32	32	24 30
Y :	44 35	34 18	22 21	10 35	47 29	31 22	40	30 32
$\mathbf{d_i}$:	-19 -4	-2 17	8 4	24 -11	-23 -15	-17 10	-8	-6 -2
$\left(d_i - \bar{d}\right)^2$:	480.92 24.30	4.93 48.02	25.70 197.96	443.94 1.144	672.36 194.04	397.20 321.4849	119.46 49.9849	79.74

$$\sum d_i = -44$$

$$\bar{d} = \frac{\sum d_i}{n} = \frac{-44}{15} = -2.93$$

$$\sum\left(d_i - \bar{d}\right)^2 = 3061.174$$

$$s = \sqrt{\frac{1}{n}\sum\left(d_i - \bar{d}\right)^2} = \sqrt{\frac{1}{15} \times 3061.174} = 14.29$$

Now, the required test statistic is

$$t = \frac{\bar{d}}{s/\sqrt{n-1}} \sim t_{n-1}$$

$$t = \frac{-2.93}{14.29/\sqrt{14}} = \frac{-2.93}{3.8198} = -0.7670$$

Critical value of t with $(n-1)=14$ d.f. at 5% l.o.s. for one til test is 1.761 (see Table 14 of Prof. T.V. Avadhani's Statistical Tables).

since the calculated value of t less than the tabulated t at 5% l.o.s. we accept H_0.

Inference:

As we are accepting H_0 we infer that there is no significant difference in food A and food B w.r.t. gain in weights of pigs.

Lesson Writer

P. NagaMani

Practical - 9

TEST FOR VARIANCES

Problem 1:

A sample of 15 observations were taken from the population with respect to the particular character of study with a population standard deviation of 40.44. Test $\sigma_2 = (40.44)^2$ at 1% l.o.s. and give your conclusion given the following observations.

43.21, 44.52, 46.31, 38.26, 43.64, 40.34, 39.99, 28.46, 43.33, 37.64, 36.54, 41.54, 36.54, 44.14, 38.64.

Aim:

To test whether the population standard deviation $\sigma = 40.44$ is significant at 1% l.o.s.

Procedure:

The test procedure is a χ^2 - test of significance of normal population variance.

Let $x_1, x_2, ..., x_n$ be a random sample drawn from the normal population with mean μ and variance σ^2.

Then $\chi^2 = \sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma}\right)^2 \sim \chi^2_{n-1}$ if population mean is not known. To test the population variance equal to a specified value, we set up the statistical hypothesis as

H_0 : There is no significant difference between population variance and sample variance.

i.e., $H_0 : \sigma^2 = \sigma_0^2 = (40 \cdot 44)^2$

Vs

H_1 : There is a significant difference between population variance and sample variance.

i.e., $H_1 : \sigma^2 \neq \sigma_0^2$.

The required test statistic is

$$\chi^2 = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2_{n-1} \text{d.f. at } \alpha \text{ \% l.o.s.}$$

If the calculated χ^2 is less than the tabulated χ^2 at $(n-1)$ d.f. for α % l.o.s. we accept H_0, otherwise reject H_0.

Calculations:

X	$\left(X-\overline{X}\right)$	$\left(X-\overline{X}\right)^2$
43.21	3.0	9.0
44.52	4.31	18.5761
46.31	6.10	37.210
38.26	-1.95	3.8025
43.64	3.43	11.7649
40.34	0.13	0.0169
39.99	-0.22	0.0484
28.46	-11.75	138.0625
43.33	3.12	9.7344
37.64	-2.57	6.6049
36.54	-3.67	13.4689
41.54	1.33	1.7689
36.54	-3.67	13.4689
44.14	3.93	15.4449
38.64	-1.57	2.4649
$\sum X = 603.10$		$\sum\left(X-\overline{X}\right)^2 = 281.4371$

$$\overline{X} = \frac{\sum X}{n} = \frac{603.10}{15} = 40.21$$

$$\chi^2 = \frac{\sum\left(X_i - \overline{X}\right)^2}{\sigma^2} = \frac{281.4371}{1635.3936} = 0.1721 \sim \chi^2_{(15-1)} \text{ d.f. at 1\% l.o.s.}$$

Tabulated value of χ^2 for 14 d.f. at 1% l.o.s. is 29.14 (see Table 15 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

Since the calculated χ^2 is less than the tabulated χ^2 at 1% l.o.s., w e accept H_0 and conclude that the population standard deviation is $\sigma = 40.44$.

Problem 2:

The two following samples are drawn from two populations. Test the equality of variances of two populations.

Sample I:	66.5,	95.2,	88.8,	23.3,	59.3,	43.3,	20.5,	83.2,	29.3,	83.6.
Sample II :	60.5,	12.1,	78.8,	29.2,	45.7,	76.4,	91.5,	36.7,	69.7,	4.7,
	15.1,	89.5,	49.1.							

Aim:

To test the equality of variances of two populations.

Procedure:

The test procedure is F - test for equality of two normal population variances.

Let $X_1, X_2, \cdots\cdots\cdots\cdots, X_n$ be a random sample drawn from a normal population with variance σ_1^2 and let $S_x^2 = \frac{1}{n_1 - 1}\Sigma\left(X_i - \overline{X}\right)^2$.

Let $Y_1, Y_2, \cdots\cdots\cdots\cdots, Y_n$ be a random sample drawn from another normal population with variance σ_2^2 and let $S_y^2 = \frac{1}{n_2 - 1}\Sigma\left(Y_j - \overline{Y}\right)^2$.

If we are interested to test the equality of two population variances, we set the hypothesis as

H_0 : The two normal population variances are equal.

i.e., $\sigma_1^2 = \sigma_2^2$

Vs

H_1 : The two normal population variances are not equal.

i.e., $\sigma_1^2 \neq \sigma_2^2$

The required test statistic is

$$F = \frac{S_x^2}{S_y^2} \sim F(n_1 - 1, n_2 - 1) \text{ d.f. under } H_0$$

If the calculated value of F is less than the tabulated value of F with $(n_1 - 1, n_2 - 1)$ d.f. at $\alpha\%$ l.o.s., we accept H_0 otherwise reject H_0.

Calculations:

X	Y	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})$	$(Y_i - \overline{Y})^2$
66.5	60.5	7.2	51.84	9.8100	96.23
95.2	12.1	35.9	1288.81	38.58	1489.18
88.8	78.8	29.5	870.25	21.10	790.17
23.3	29.2	35.62	1296.00	21.48	461.82
59.3	45.7	0	0	4.98	24.90
43.3	76.4	16.0	256.0	25.71	661.00
20.5	91.5	38.8	1505.44	40.81	1665.45
83.2	36.7	23.9	571.21	13.98	195.72
29.3	69.7	30.0	900.00	19.00	361.38
83.6	4.7	24.3	590.49	45.98	2115.08
	15.1			35.58	1266.64
	89.5			38.81	1506.21
	49.1			1.58	2.52
$\sum X =$ 593	$\sum Y =$ 659		$\sum (X - \overline{X})^2$ = 7330.04		$\sum (Y - \overline{Y})^2$ =10636.30

$$\overline{X} = \frac{\sum X}{n_1} = \frac{593}{10} = 59.3$$

$$\overline{Y} = \frac{\sum Y}{n_2} = \frac{659}{13} = 50.69$$

$$S_X^2 = \frac{1}{n_1 - 1} \Sigma\left(X_i - \overline{X}\right)^2$$

$$= \frac{1}{9} \times 7330.04 = 814.44$$

$$S_Y^2 = \frac{1}{n_2 - 1} \Sigma\left(Y_j - \overline{Y}\right)^2$$

$$= \frac{1}{12} \times 10636.30 = 886.35$$

$$F = \frac{S_X^2}{S_Y^2} \sim F\left(n_1 - 1,\ n_2 - 1\right)$$

$$F = \frac{886.35}{814.44} \sim F(9,12)$$

$$F = 1.08$$

Tabulated value of F for (9, 12) d.f. at 5% l.o.s. is 2.80 (see Table 16 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

Since, the calculated value of F is less than the tabulated value of F, we accept H_0 and conclude that the variances of two normal populations are equal.

Lesson Writer
P. NagaMani

Practical - 10

χ^2 TEST FOR GOODNESS OF FIT

Problem 1:

A coin is tossed 100 times with the following results:

x represents the number of heads appeared and f represents the corresponding frequency.

x :	0	1	2	3	4	5	6	7	8
f :	4	12	24	32	14	8	4	1	1

Test whether the coin is unbiased at 5% level of significance.

Aim:

To test the unbiasedness of the coin at 5% level of significance.

Procedure:

The test procedure for testing the unbiasedness of the coin is χ^2 - test for goodness of fit.

Let us suppose that $O_1, O_2, \cdots\cdots\cdots\cdots\cdots, O_n$ are the observed or experimental frequencies and $e_1, e_2, \cdots\cdots\cdots e_n$ are the corresponding expected frequencies. To test the fitness between the observed and expected frequencies we setup the hypothesis as

H_0 : The fit between the observed and expected frequencies is good.

i.e., The coin is unbiased

Vs

H_1 : The fit between the observed and expected frequencies is not good.

i.e., The coin is not unbiased

The required test statistic to test H_0 is

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - e_i)^2}{e_i}\right] \sim \chi^2_{n-1} \text{ d.f. at } \alpha\ \% \text{ l.o.s.}$$

If the calculated value of χ^2 is less than the tabulated value of χ^2 at (n - 1) d.f. at α% l.o.s., we accept H_o, otherwise reject H_o.

In this problem, the expected frequencies are calculated by using the binomial distribution whose p.d.f. is given by

$$p(x) = {}^{n}C_{x}\, p^{x} q^{n-x} \;\; ; \; x = 0,1,2,.......,n$$

Here, $p = \frac{1}{2},\; q = \frac{1}{2};$

$n = 8.$

Now calculate $p(x)$ for all $x = 0,1,2,.........,8$.

Expected frequencies are obtained by multiplying each of these probabilities by $N = \sum f$ and rounding off to its nearest integer.

Calculations:

Calculation of expected frequencies.

$$p(0) = {}^{8}C_{0}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{8-0} = 0.00390$$

$$E(o) = 100 \times 0.00390 = 0.390 \simeq 0$$

$$p(1) = {}^{8}C_{1}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{7} = 0.03125$$

$$E(1) = 100 \times 0.03125 = 3.125 \simeq 3$$

$$p(2) = {}^{8}C_{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{6} = 0.1093$$

$$E(2) = 100 \times 0.1093 = 10.9375 \simeq 11$$

$$p(3) = {}^{8}C_{3}\left(\frac{1}{2}\right)^{3}\left(\frac{1}{2}\right)^{5} = 0.21875$$

$$E(3) = 100 \times 0.21875 = 21.875 \simeq 22$$

$$p(4) = {}^{8}C_{4}\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right)^{4} = 0.2734$$

$$E(4) = 100 \times 0.2734 = 27.34 \simeq 27$$

$$p(5) = {}^{8}C_{5}\left(\frac{1}{2}\right)^{5}\left(\frac{1}{2}\right)^{3} = 0.21875$$

$$E(5) = 100 \times 0.21875 = 21.875 \simeq 22$$

$$p(6) = {}^8C_6\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^2 = 0.10937$$

$$E(6) = 0.10937\times100 = 10.9375 \simeq 11$$

$$p(7) = {}^8C_7\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right) = 0.03125$$

$$E(7) = 100\times0.03125 = 3.125 \simeq 3$$

$$p(8) = {}^8C_8\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^0 = 0.00390$$

$$E(8) = 100\times0.00390 = 0.3906 \simeq 0$$

Observed Frequencies	**Expected Frequencies**	$(O_i - e_i)^2$	$(O_i - e_i)^2/e_i$
4	0	16	0
12	3	81	27
24	11	169	15.36
32	22	100	4.54
14	27	169	6.25
8	22	196	8.90
4	11	49	4.45
1	3	4	1.33
1	0	1	0.00
			$X^2 = 67.83$

Tabulated value of χ^2 for 7 d.f. at 5% l.o.s. is 14.067 (see Table 15 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

Since, the calculated value of χ^2 is greater than the tabulated value of χ^2 for 7 d.f. at 5% l.o.s., we reject H_0 and conclude that the binomial fit is not good for the given data. In other words the coin is a biased coin.

Lesson Writer
P. NagaMani

Practical - 11

χ^2 TEST FOR INDEPENDENCE OF ATTRIBUTES

Problem:

The following data gives the classification of college students with respect to their intelligence and economic conditions.

Economic	Intelligence			
Conditions	Excellent	Good	Medium	Dull
High	24	36	42	53
Good	32	44	62	18
Bad	48	38	15	10
Worse	40	30	15	10

Test whether there is association between intelligence and economic conditions.

Aim:

To test whether there is association between intelligence and economic conditions of the students.

Procedure:

Let A denote the economic conditions and B denote the intelligence of the students. Attribute A is divided into four classes A_1, A_2, A_3 and A_4 and atttribute B is divided into four classes B_1, B_2, B_3 and B_4.

We have the observed frequencies (A_iB_j) representing the number of individuals possessing the attributes A_i and B_j simulataneously $\forall\ i = 1,2,3,4\ \&\ j = 1,2,3,4$.

(A_i) indicates the number of individuals possessing attribute A_i.

(B_j) indicates the number of individuals possessing the attribute B_j.

$$N = \sum\sum(A_iB_j)$$

Expected frequency for each observed frequency is calculated by

$$\left(A_iB_j\right)_0 = \frac{\left(A_i\right)\left(B_j\right)}{N} \ \forall\ i = 1,2,3,4\ ;\ j = 1,2,3,4\,.$$

To test the independence of two attributes, we setup the hypothesis as

H_0 : The two attributes A and B are independent.

Vs

H_1 : The two attributes A and B are associated.

The required test statistic to test the above hypothesis is

$$\chi^2 = \sum_i \sum_j \left\{ \frac{\left[\left(A_iB_j\right) - \left(A_iB_j\right)_0\right]^2}{\left(A_iB_j\right)_0} \right\} \sim \chi^2_{(r-1)(s-1)} \text{ d.f. at } \alpha\% \text{ l.o.s.}$$

If the calculated value of χ^2 less than the tabulated value of χ^2 at $(r-1)(s-1)$ d.f. at α % l.o.s., we accept H_0, otherwise reject H_0.

Calculations:

Economic Conditions	**Intelligence**				
	A_1	A_2	A_3	A_4	
B_1	24	36	42	53	155
B_2	32	44	62	18	156
B_3	48	38	15	10	111
B_4	40	30	15	10	95
	144	148	134	91	517

(A_iB_j)	$(A_iB_j)_0$	$\left[(A_iB_j)-(A_iB_j)_0\right]^2$	$\left[(A_iB_j)-(A_iB_j)_0\right]^2 / (A_iB_j)_0$
24	43	361	8.3953
32	43	121	2.8139
48	31	289	9.3225
40	26	196	7.5384
36	44	64	1.4545
44	45	1	0.0222
38	32	36	1.1250
30	27	9	0.3333
42	40	4	0.1000
62	40	484	12.100
15	29	196	6.7586
15	25	100	4.0000
53	27	676	25.0370
18	27	81	3.0000
10	20	100	5.0000
10	17	49	2.8823
			χ^2 = 89.8827

Tabulated value of χ^2 at $(r-1)(s-1)=(4-1)(4-1)=9$ d.f. at 5% l.o.s. is 16.919 (see Table 15 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

Since, the calculated value of χ^2 is greater than the tabulated value for 9 d.f. at 5% l.o.s., we reject H_0 and conclude that the two attributes intelligence and economic conditions are associated.

Lesson Writer
P. NagaMani

Practical - 12

NP TESTS FOR ONE - SAMPLE

Problem 1:

The following observations are drawn at random

104, 146, 132, 124, 111, 146, 123, 141, 91, 98, 75, 86, 99, 108, 107, 93, 86, 24, 48, 100, 103.

Test the randomness of the data using run test.

Aim:

To test the randomness of the observations using run test.

Procedure:

Let $X_1, X_2, ..., X_n$ be the set of observations taken from the population. If we are interested to test the randomness of the observations, we set the hypothesis.

H_0 : The sample observations are drawn at random

Vs

H_1 : The sample observations are not drawn at random

The test procedure is to find median M for the given observations. Indicate the observations by letter A which are less than M and the observations by letter B which are greater than M, getting a sequence of letters. Now, count the number of runs 'r'. If $r_1 < r < r_2$, we conclude that the observations are drawn at random, otherwise reject H_0. r_1 and r_2 are obtained from run tables at n_1 (number of A's) and n_2 (number of B's) (for critical values of 'r' see the statistical tables by Prof. T.V. Avadhani of Andhra University).

For large samples usually n_1 or $n_2 > 20$, $r \sim N[E(r), V(r)]$

where,

$$E(r) = \frac{2n_1n_2}{n_1+n_2} + 1 \quad \text{and} \quad V(r) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

The required test statistic is

$$Z = \frac{r - E(r)}{S.E.(r)} \sim N(0,1)$$

If calculated value of $|Z| <$ tabulated Z at $\alpha\%$ l.o.s. We accept H_0, otherwise we reject H_0.

Calculations:

First arrange the observations in ascending order

24, 48, 75, 86, 86, 91, 93, 98, 99, 100, 103, 104, 107, 108, 111, 123, 124, 132, 141, 146, 146.

Median = 103

Consider the given observations in their natural order

B	B	B	B	B	B	B	B	A	A	A	A	A	B	B
104,	146,	132,	124,	111,	146,	123,	141,	91,	98,	75,	86,	99,	108,	107,

A	A	A	A	A	
93,	86,	24,	48,	100,	103

n_1 = no of A's = 10

n_2 = no of B's = 10

r = no of runs = 4

r_1 = 6

r_2 = 16

Since $r < r_1 < r_2$, we reject H_0.

(see Tables 19(a) and 19(b) of Prof. T.V. Avadhani's Statistical Tables).

Inference:

As we are rejecting H_0 we infer that the observations are not drawn at random.

Problem 2:

A random sample of 22 observations were drawn from a population with population median 95. The observations are as follows:

24, 48, 93, 88, 115, 107, 82, 97, 103, 86, 113, 107, 112, 98, 90, 93, 99, 103, 104, 100, 101, 96.

Test the population median is 95 using sign test.

Aim:

To test the population median M = 95 using sign test.

Procedure:

The test procedure to test the hypothesis is one sample sign test.

Let $x_1, x_2, ..., x_n$ be a random sample drawn from the unknown continuous population with median M. If we are interested to test the population median, we set up the hypothesis as

H_0 : The sample has been drawn from the population with median $M = M_0 = 95$

Vs

H_1 : The sample has not been drawn from the population with median $M \neq 95$.

First arrange the observations in ascending order i.e., $x_1 < x_2 < .... < x_n$.

Let $d_i = (x_i - M_0)$ under H_0. Consider the signs of deviations d_i excluding 'o' deviations. We have $P_r\{d_i < 0\} = P_r\{d_i > 0\} = 1/2$.

[$\because$ Median is the value which divides the entire data into two equal halfs].

Let x be number of +ve or -ve signs of d_i out of 'n' signs which follows Binomial with the parameters $n, p = \frac{1}{2}$, then $p(x) = {}^nC_x \, p^x \, (1-p)^{n-x} = {}^nC_x \left(1/2\right)^n$.

The test criterian is $2F_x(U) = 2P(x \le u) = 2 \cdot \sum_{x=0}^{U} p(x)$

$$= 2 \sum_{x=0}^{U} {}^nC_x \left(\frac{1}{2}\right)^n$$

$$= 2\left(\frac{1}{2}\right)^n \sum_{x=0}^{U} {}^nC_x \sim \alpha \text{ l.o.s.,}$$

where U is the total of +ve or -ve signs whichever is least.

If the calculated value of $2F_x(U) > \alpha$, we accept H_0, otherwise reject H_0.

Particularly, if the sample is large i.e., $n \ge 25$, we use standard normal test criterian to test the hypothesis.

Since $x \sim B\left(n, p = \frac{1}{2}\right)$

$$Z = \frac{x - E(x)}{S.E.(x)} = \frac{U - n/2}{\sqrt{n/4}} = \frac{2u - n}{\sqrt{n}} \sim N(0,1) \text{ under } H_0.$$

If the calculated value of Z is less than the tabulated Z at α % l.o.s. we accept H_0, otherwise reject H_0.

Calculations:

Let us arrange the observations in the ascending order.

24, 48, 82, 86, 88, 90, 93, 93, 96, 97, 98, 99, 100, 101, 103, 103, 104, 107, 107, 112, 115, 115.

Given $M_0 = 95$ and $d_i = x_i - M_0$

$\therefore$ No of -ve signs of $d_i = 8$

No of +ve signs of $d_i = 14$

Let U = 8.

The test criterian is

$$2F = 2 \sum_{x=0}^{U} {}^{n}C_x \left(\frac{1}{2}\right)^n$$

$$= \frac{1}{2^{n-1}} \sum_{x=0}^{8} \left({}^{22}C_x\right)$$

$$= \frac{1}{2^{21}} [600370]$$

= 0.2863

The critical value to be compared is 5% = 0.05

Inference:

Since the calculated value is greater than the considered level of significance we accept H_0 and infer that the population median is M = 95.

Problem 3:

A random sample of 20 observations from a population are

104, 96, 101, 100, 103, 99, 93, 98, 90, 112, 104, 110, 86, 103, 97, 82, 115, 107, 88, 93.

Test the population mean is 100 using wilcoxon signed rank test.

Aim:

To test the population median is 100 using Wilcoxon signed rank test.

Procedure:

Let $X_1, X_2, ..., X_n$ be a random sample drawn from an unknown continuous population with median M. If we are interested to test the population median $M = M_0$, we setup the hypothesis as:

H_0 : The given sample has been drawn from the population with the median $M = M_0 = 100$.

Vs

H_1 : The given sample has been drawn from the population with the median $M \neq 100$.

The test procedure is to first arrange the data in ascending order. Then calculate $d_i = (X_i - M_0)$ which may have both +ve's and -ve's. Discarding the sign of d_i assign ranks to $d_i \ \forall \ i = 1, 2, ..., n$. Let t^+ represents the sum of ranks of d_i corresponding to +ve signs and t^- represent, the sum of ranks of d_i corresponding to -ve signs.

Now, the corresponding test criteria is $t = \text{Min}\left(t^+, t^-\right)$.

If $t > t_\alpha$, we accept H_0 and conclude that the sample has been drawn from the population with the median $M = M_0$. Here t_α is the critical value obtained from Wilcoxon signed ranks table corresponding to the sample size 'n' at $\alpha\%$ l.o.s.

Calculations:

Let us first arrange the data in ascending order

82, 86, 88, 90, 93, 93, 96, 97, 98, 99, 100, 101, 103, 103, 104, 104, 107, 110, 112, 115.

Given $M_0 = 100$,

$d_i = (X_i - M_0)$

d_i : -18, -14, -12, -10, -7, -7, -4, -3, -2, -1, 0, 1, 3, 3, 4, 4, 7, 10, 12, 15.

Ranks

to

$|d_i|$: 19, 17, 15.5, 13.5, 10, 10, 8, 5, 3, 1.5, 1.5, 5, 5, 8, 8, 10, 13.5, 15.5, 18.

t^+ is sum of ranks of +ve d_i's = 84.5

t^- is sum of ranks of -ve d_is = 102.5

t = min $\left(t^+, t^-\right) = 84.5$

As one of $d_i = 0$, $n = 20 - 1 = 19$.

Critical value of t_α at 5% l.o.s. for n = 19 from Wicoxon signed rank tables is 46 (see Table 22 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

Since, $t > t_\alpha$, we accept H_0 and conclude that the population median is 100.

Lesson Writer

P. NagaMani

Practical - 13

NP TESTS FOR TWO RELATED SAMPLES

Problem 1:

18 people were selected for treatment of certain disease. The following data reveals the grades of these people before and after they were given treatment. Test whether the condition of people irrproved after the treatment or not using sign test.

Grades Before Treatment:

1 3 5 2 6 2 2 3 2 3 3 3 3 3 3 3 2 2

Grades After Treatment:

5 5 5 5 4 5 5 5 1 3 5 2 3 5 5 4 4 2

Aim:

To test whether the condition of the people were improved after the treatment using sign test.

Procedure:

Let $X_1, X_2, \ldots\ldots\ldots\ldots, X_n$ be a sample of size 'n' from an unknown continuous population with its form $f(x)$.

Let $Y_1, Y_2, \ldots\ldots\ldots, Y_n$ be another random sample of size 'n' from another unknown continuous population with its form $f(y)$. If we want to test the significant difference between the two populations, the hypothesis for testing is

H_0 : The condition of the people is same before and after the treatment.

i.e., $H_0 : f(x) = f(y)$

Vs

H_1 : The condition of the people is not same before and after the treatment.

i.e., $H_1 : f(x) \neq f(y)$

The test procedure is to observe the signs of deviations $d_i = (x_i - y_i)$. Now exclude 'O' deviations and take 'U' as number of +ve signs.

Get the values of u_1 and u_2 at $\alpha/2$ and $1-\alpha/2$ l.o.s. respectively for the sample size 'n' obtained from binomial cumulative probabilities.

If $u_1 < u < u_2$, we accept H_0 and conclude that there is no significant difference between the two populations. Otherwise reject H_0 and say that there is a significant difference between the two populations.

Calculations:

x	:	1	3	5	2	6	2	2	3	2	3
		3	3	3	3	3	3	2	2		
y	:	5	5	5	5	4	5	5	5	1	3
		5	2	3	5	5	4	4	2		
$d_{(x_i - y_i)}$	:	-	-	0	-	+	-	-	-	+	0
		-	+	0	-	-	-	-	0		

Here, the sampel size is 18. As we have 4 zero deviations, the size of the sample reduces by 4.

i.e., $n = 18 - 4 = 14$.

If $\alpha = 5\%$ the value of $u_1 = 2$ at $\alpha/2 = 0.025$ for n = 14, and value of $u_2 = 10$ at $1 - \alpha/2 = 0.975$ for n = 14 (see Table 24 of Prof. T.V. Avadhani's Statistical Tables).

Here no of +ves = 3.

no of -ves = 11 $\therefore\ u = 3$

Here $u_1 < u < u_2$; i.e., $2 < 3 < 10$ we accept H_0.

Inference:

We infer that there is no improvement in the 18 people before and after the treatment.

Problem 2:

In a college a random sample of 26 students were taken and their marks in prefinal and final exams in a subject is as follows. Test whether the prefinal and final exam marks have the same distribution function using Wilcoxon signed rank test.

Prefinal: 83, 75, 34, 70, 83, 67, 20, 75, 26, 87, 23, 21, 75, 33, 44, 9, 65, 44, 64, 25, 83, 75, 34, 70, 83, 67.

Final: 33, 98, 70, 64, 99, 60, 68, 58, 62, 76, 26, 41, 03, 26, 26, 55, 62, 61, 82, 19, 33, 98, 70, 64, 99, 60.

Aim:

Test whether the prefinal and final exam mraks have the same distribution function using Wilcoxon signed rank test.

Procedure:

Let $X_1, X_2, \cdots\cdots\cdots, X_n$ be a random sample drawn from unknown continuous population with its form $f(x)$.

Let $Y_1, Y_2, \cdots\cdots\cdots, Y_n$ be another random sample drawn from other continuous population with its form $f(y)$.

If we want to test the significant difference between the two populations we setup the hypothesis as

H_0 : There is no significant difference between the two populations. i.e., The prefinal and final exams marks have the same distribution function.

i.e., $H_0 : f(x) = f(y)$

Vs

H_1 : There is a significant difference between the two populations. i.e., the prefinal and final exams marks do not have the same distribution function.

i.e., $H_1 : f(x) \neq f(y)$

The test procedure is to give ranks for $|d_i|$, where $d_i = x_i - y_i \ \forall i = 1, 2, ..., n$. Let t^+ be the sum of ranks of $|d_i|$ corresponding to +ve signs and let t^- be the sum of ranks of $|d_i|$ corresponding to -ve signs. The test criteria is $t = \min\left(t^+, t^-\right)$.

Since, sample size n > 25 $t \sim N\left[E(t), V(t)\right]$

Where $E(t) = \frac{n(n+1)}{4}$ and $V(t) = \frac{n(n+1)(2n+1)}{24}$

Test criteria is

$$Z = \frac{t - E(t)}{SE(t)} \sim N(0,1)$$

If the calculated value of $|Z|$ < the tabulated value of Z at α % l.o.s. we accept H_0. Otherwise reject H_0.

Calculations:

Prefinal X_i	**Final** Y_i	$\lvert d_i \rvert$	**Ranks of** $\lvert d_i \rvert$
83	33	50	24.5
75	98	23	17.5
34	70	36	20
70	64	6	4
83	99	16	10.5
67	60	7	7
20	68	48	23
75	58	17	12.5
26	62	36	20
87	76	11	9
23	26	3	1.5
21	41	20	16
75	03	72	26
33	26	7	7
44	26	18	14.5
9	55	46	22
65	62	3	1.5
44	61	17	12.5
64	82	18	14.5
25	19	6	4
83	33	50	24.5
75	98	23	17.5
34	70	36	20
70	64	6	4
83	99	16	10.5
67	60	7	7

$$t = \min\left(t^{+},\ t^{-}\right) = \min\left(145.5,\ 205.5\right) = 145.5$$

$$E\left(t\right) = \frac{n\left(n+1\right)}{4} = \frac{26\times 27}{4} = 175.5$$

$$V\left(t\right) = \frac{n\left(n+1\right)\left(2n+1\right)}{24} = \frac{26\times 27\times 53}{24} = 1550.25$$

$$\text{S.E.}\left(t\right) = \sqrt{1550.25} = 39.3732$$

$$Z = \frac{t - E\left(t\right)}{\text{S.E.}\left(t\right)} \sim N\left(0,1\right)$$

$$Z = \frac{145.5 - 175.5}{39.3732}$$

$$Z = -0.7619$$

$$\left|Z\right| = 0.7619$$

Tabulated value of Z at 5% l.o.s. for two - tailed test is 1.96 (see Table 13 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

since, the calculated value of $\left|Z\right|$ is less than the tabulated value of Z at 5% l.o.s. we accept H_0 and infer that the marks of final and prefinal exams have the same distribution.

Lesson Writer

P. NagaMani

Practical - 14

NP TESTS FOR TWO INDEPENDENT SAMPLES

Problem 1:

The following is the data related to temperature in degress in two places.

Place I : 24, 32, 28, 40, 22, 24, 26, 32, 36, 38, 39, 40, 21, 23, 26, 29, 20, 22, 26, 27, 19.

Place II : 16, 18, 19, 24, 23, 22, 26, 24, 28, 32, 40, 36, 30, 31, 34, 26, 35, 36, 38, 40, 41.

Test whether the two places have the same temperature using run test.

Aim:

To test whether the two places have the same temperature using run test.

Procedure:

Let $X_1, X_2, ..., X_{n_1}$ be a random sample of size n_1 drawn from unknown continuous population with the density $f(x)$ and let $Y_1, Y_2, ..., Y_{n_2}$ be another sample of size n_2 drawn from another continuous population which is unknown with the density $f(y)$.

If we want to test the significant difference between the two populations we set up the hypothesis as

H_0 : There is no significant difference in the temperature of two places.

i.e., $f(x) = f(y)$

Vs

H_1 : There is significant difference in the temperatures of two places.

i.e., $f(x) \neq f(y)$.

The test procedure is to combine the two samples as to make a single sample of size $n = n_1 + n_2$ and then arrange the observations in ascending order, then represent the observations by letter A if it belongs to sample - I and the observations by letter B if it belongs to sample II for the ascending order of the sample of size 'n'. Then we get the sequence of two letters A and B, and count the no of runs 'r'.

For large sample n_1 or n_2, i.e., ≥ 20 the test statistic is

$$Z = \frac{r - E(r)}{S.E.(r)} \sim N(0,1)$$

$$E(r)=\frac{2n_1n_2}{n_1+n_2}+1$$

$$V(r)=\frac{2n_1n_2\left(2n_1n_2-n_1-n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2-1\right)}$$

If calculated Z is less than the tabulated value of Z at α % l.o.s. we accept H_0 and conclide that there is no significant difference between the two populations, otherwise reject H_0.

Calculations:

First arrange the data in ascending order combining both the samples.

Let A stands for place I

B stands for place II

$$\begin{matrix}\overline{B\quad B\quad B} & \overline{A\quad A\quad A\quad A\quad A} & \overline{B} & \overline{A} & \overline{B} & \overline{A\quad A} & \overline{B\quad B}\\ 16\quad 18\quad 19 & 19\quad 20\quad 21\quad 22\quad 22 & 22 & 23 & 23 & 24\quad 24 & 24\quad 24\end{matrix}$$

$$\begin{matrix}\overline{A\quad A\quad A} & \overline{B\quad B} & \overline{A\quad A} & \overline{B} & \overline{A} & \overline{B\quad B} & \overline{A\quad A} & \overline{B\quad B\quad B}\\ 26\quad 26\quad 26 & 26\quad 26 & 27\quad 28 & 28 & 29 & 30\quad 31 & 32\quad 32 & 32\quad 34\quad 35\end{matrix}$$

$$\begin{matrix}\overline{A} & \overline{B\quad B} & \overline{A} & \overline{B} & \overline{A\quad A\quad A} & \overline{B\quad B\quad B}\\ 36 & 36\quad 36 & 38 & 38 & 39\quad 40\quad 40 & 40\quad 40\quad 41\end{matrix}$$

$$r=21$$

$$Z=\frac{r-E(r)}{S.E.(r)}\sim N(0,1)$$

$$E(r)=\frac{2n_1n_2}{n_1+n_2}+1=\frac{2\times21\times21}{21+21}+1=\frac{882}{42}+1=22$$

$$V(r)=\frac{2n_1n_2\left(2n_1n_2-n_1-n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2-1\right)}$$

$$=\frac{2\times21\times21\left(2\times21\times21-21-21\right)}{\left(21+21\right)^2\left(21+21-1\right)}$$

$$= \frac{882 \times 840}{1764 \times 41}$$

$$= \frac{740880}{72324} = 10.2439$$

$$\text{S.E.}(r) = 3.20$$

$$Z = \frac{r - E(r)}{\text{S.E.}(r)} \sim N(0,1)$$

$$Z = \frac{21 - 22}{3.20} = -0.3125$$

$$|Z| = 0.3125$$

Tabulated value of Z at 5% l.o.s. for two tailed test is 1.96 (see Table 13 of Prof. T.V. Avadhani's Statistical Table).

Inference:

Since, calculated $|Z| <$ tabulated Z at 5% l.o.s. we accept H_0 and infer that there is no significant difference in the temperatures of two places.

Problem 2:

The following are the scores obtained for 8 clearks in a parliament and 9 clerks in secretariat in an eligibility test for their performance scores of clerks.

Scroes of Clerks parliament :	40, 35, 52, 60, 46, 55, 62, 61
Scores of Clerks in Secretariat :	47, 56, 42, 57, 50, 67, 62, 61, 58

Test whether the performance of clerks in two offices are equal in the test for promotion using Wilcoxon Mann-Whitney u - test.

Aim:

To test whether the performance of clerks in two offices are equal in the test for promotion using Wilcoxon - Mann Whitney U - test.

Procedure:

Let $X_1, X_2, \ldots\ldots\ldots, X_n$ and $Y_1, Y_2, \ldots\ldots\ldots\ldots\ldots\ldots, Y_{n_2}$ be two independent random samples of sizes n_1 and n_2 drawn from two independent populations with p.d.f. $f(x)$ and $f(y)$ respectively.

If we are interested to test the equality of two populations. We setup the statistical hypothesis as

H_0 : The two populations are identical.

i.e., $f(x)=f(y)$ i.e., there is no significant difference in the performance of clerks in two offices.

Vs

H_1 : The two populations are not identical.

i.e., $f(x)\neq f(y)$ i.e., There is a significant difference in the performance of clerks in two offices.

The test procedure is based on the pattern of the X's and Y's in the combined ordered sample. Let T denote the sum of ranks of Y's in the combined ordered sample. The test statistic U defined in terms of T as follows:

$$U=\left(n_1n_2+\frac{n_2(n_2+1)}{2}\right)-T$$

For n_1 or $n_2>8$ standard normal test criteria to test H_0 is

$$Z=\frac{U-E(U)}{S.E.(U)}\sim N(0,1)$$

where $E(U)=\frac{n_1n_2}{2}$ and $V(U)=\frac{n_1n_2(n_1+n_2+1)}{12}$.

If the calculated value of $|Z|<$ tabulated value of Z at α % l.o.s. we accept H_0 and conclude that the two populations are identical otherwise reject H_0.

Calculations:

Let us combine both the samples and arrange in the ascending order.

	X	X	Y	X	Y	Y	X	X	Y	Y	Y	X	X	Y	X	Y	Y
	35	40	42	46	47	50	52	55	56	57	58	60	61	61	62	62	67
Ranks :	1	2	3	4	5	6	7	8	9	10	11	12	13.5	13.5	15.5	15.5	17

T = sum of ranks of Y's = 90

Since $n_2>8$ the test criteria is standard normal criteria given by

$$Z = \frac{U - E(U)}{S.E.(U)} \sim N(0,1)$$

where $E(U) = \frac{n_1 n_2}{2} = \frac{8 \times 9}{2} = 36$

$$V(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{8 \times 9(8+9+1)}{12} = 108$$

$$S.E.(U) = 10.3923$$

$$U = \left[n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} \right] - T$$

$$U = \left[8 \times 9 + \frac{9(9+1)}{2} \right] - 90$$

$$U = [72 + 45] - 90 = 27$$

$$Z = \frac{27 - 36}{10.3923} = -0.866$$

$$|Z| = 0.866$$

Tabulated value of Z at 5% l.o.s. for two tailed test is 1.96 (see Table 13 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

Since calculated $|Z|$ is less than the tabulated value of Z at 5% l.o.s. we accept H_0 and infer that the performance of clerks in two offices are equal in the test for promotion.

Problem 3:

The school children taking coaching in two private schools secured the following scores out of 100.

School 1:	33, 38, 39, 48, 58, 70, 61, 41, 45, 49
School 2:	32, 15, 87, 32, 22, 63, 56, 57, 44

Test the hypothesis that the students studying in two private schools have identical distribution of marks by applying Median test at 1% l.o.s.

Aim:

To test the hypothesis that the students in two private schools have identical distribution of marks by applying median test at 1% l.o.s.

Procedure:

Let $X_1, X_2, ..., X_{n_1}$ and $Y_1, Y_2, ..., Y_{n_2}$ be two random samples of sizes n_1 and n_2 drawn from two populations with the functions $F(x)$ and $F(y)$ respectively. If we are interested to test the equality of two populations with the same median, we setup the hypothesis as

$$H_0 : F_x(x) = F_y(x) \ \forall \ x$$

Vs

$$H_1 : F_x(x) = F_y(x-\delta) \ \forall \ x \ \ \& \ \delta \neq 0$$

where δ is the shift in the location parameter which is the median in the test. In this problem

H_0 : The students studying in two private schools have identical distribution of marks.

Vs

H_1 : The students studying in two private schools do not have identical distribution of marks.

The test procedure for median test is to combine the samples and arrange them in order. Then find the median $'\theta'$ for the ordered sample. Count number of X's and Y's on the left of $'\theta'$. Let 'u' represents number of X's to the left of θ and v represents number of Y's to the left of θ. The test is based on u, the number of observations which are less than θ in the combined sample.

For $n \geq 10$; i.e., $n_1 + n_2 \geq 10$, the test is a standard test.

$$Z = \frac{U - E(U)}{S.E.(U)} \sim N(0,1)$$

$$E(U) = \frac{n_1 t}{n}; \ V(U) = \frac{n_1 n_2 (n-t)}{n^2 (n-1)}$$

where

n_1 = no of observations in sample I

n_2 = no of observations in sample II

$n = n_1 + n_2$

$$t = \frac{n}{2}$$

If the calculated $|Z|$ is less than the tabulated Z at $\alpha\%$ l.o.s. we accept H_0 and conclude that the two populations have the same median.

Calculations:

Combined sample in the ascending order is

y	y	y	y	x	x	x	x	y	x	x	x	y	y	x	x	y	x	y
15	22	32	32	33	38	39	41	44	45	48	49	56	57	58	61	63	70	87

The median to the data is 45.

$$u = 4 \quad ; v = 5, \quad n = \underset{10}{n_1} \underset{+}{+} \underset{9}{n_2} = 19, \; t = \frac{19}{2} = 9.5 \simeq 10$$

Since, n > 10 the test is a standard normal test.

$$Z = \frac{U - E(U)}{S.E.(U)} \sim N(0,1)$$

$$E(U) = \frac{n_1 t}{n} = \frac{10 \times 10}{19} = 5 \cdot 263$$

$$V(U) = \frac{n_1 n_2 (n - t)}{n^2 (n-1)} = \frac{10 \times 9(19 - 10)}{(19)^2 (18)} = \frac{810}{6498} = 0.1246$$

$$S.E.(U) = 0.3530$$

$$Z = \frac{4 - 5.263}{0.3530} = \frac{-1.263}{0.3530} = -3.58; \quad |Z| = 3.58$$

Tabulated value of $|Z|$ at 5% l.o.s. is 1.96 (see Table 13 of Prof. T.V. Avadhani's Statistical Tables).

Inference:

Since, the calculated $|Z| >$ tabulated Z at 5% l.o.s. we reject H_0 and infer that the students studying in two private schools do not have identical distribution of marks.

Lesson Writer
P. NagaMani