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

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – I – Statistical Inference and Stochastic Process– SMT5208

THEORY OF ESTIMATION

Properties of estimates; asymptotically most efficient estimates; Likelihood function; Cramer-Rao inequality; Rao-Black-Well`s theorem; Properties of Maximum likelihood estimates. Problems related to Maximum likelihood estimates.

THE THEORY OF ESTIMATION:

The theory of estimation is the procedure of estimating an unknown parameter on the basis of a sample is the following: On the basis of a sample we determine the value u of a certain statistic U whose distribution depends upon this parameter and we take this value u as an estimate of the unknown parameter. Both the statistic U and its observed value u are called the estimate of the unknown parameter.

UNBIASEDNESS:

If $E(T_n) = \gamma(\theta)$ for all $\theta \epsilon \Theta$. Then T_n is said to be an unbiased estimator of $\gamma(\theta)$.

EXAMPLE:

Let $x_1, x_2, \ldots, x_n$ be a random sample of X having unknown mean μ. Show that the estimator of μ defined by

$M = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an unbiased estimator of μ

$$E(M) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} E(x_i)\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu$$

$$= \mu$$

Therefore, M is an unbiased estimator of μ.

ASYMPTOTICALLY MOST EFFICIENT ESTIMATE:

Estimates which are not most efficient but their efficiency satisfies the condition

$$\lim_{n\to\infty} e = 1,$$

and they are at least asymptotically unbiased. Such estimates are called asymptotically most efficient estimates. Since the efficiency e may converge to a limit $e_0 \neq 1$ as $n \to \infty$. The limit value of the number e is called the asymptotic efficiency of the estimate.

EXAMPLE :

The characteristic X of elements of a population of the normal distribution N(m;σ). We want to estimate from simple samples of n elements the variance σ^2 and we take as an estimate the statistic

$$U = \frac{n}{n-1} S^2 = \frac{\sum_{k=1}^{n}(X_k - \overline{X})^2}{n-1}$$

The estimate U is unbiased and regular in an arbitrary interval. This estimate is not sufficient and hence is not the most efficient.

The efficiency of this estimates equals

$$e = \frac{n-1}{n} < 1$$

It is easy to see that $\lim_{n\to\infty} e = 1$. Hence the considered estimate is asymptotically most efficient.

LIKELIHOOD FUNCTION:

$x_1, x_2, \ldots, x_n$ is a random sample of size n with density function f(x,θ) then the likelihood function is defined as

$L = f(x_1,\theta).f(x_2,\theta)\ldots f(x_n,\theta)$

$= \prod_{i=1}^{n} f(x_i, \theta).$

LIKELIHOOD FUNCTION OF THE DISCRETE MARKOV PROCESS:

The product,

$$L = C_{j_0}(t_0, \lambda_1, \lambda_2, \ldots, \lambda_m) \prod_{k=0}^{n-1} P_{j_k j_{k+1}}(t_k, t_{k+1}, \lambda_1, \lambda_2, \ldots, \lambda_m),$$

where $C_{j_0}(t_0, \lambda_1, \lambda_2, \ldots, \lambda_m) = P(X_{t_0} = j)$ is the absolute probability function of the random variable X_{t_0}, is called the likelihood function of the discrete Markov process.

LIKELIHOOD FUNCTION OF THE CONTINUOUS MARKOV PROCESS:

The product,

$$L = f(t_0, x_0, \lambda_1, \lambda_2, \ldots, \lambda_m) \prod_{k=0}^{n-1} f(t_k, x_k, t_{k+1}, x_{k+1}, \lambda_1, \lambda_2, \ldots, \lambda_m),$$

where $f(t_0, x_0, \lambda_1, \lambda_2, \ldots, \lambda_m)$ is the density of the absolute distribution of the random variable X_{t_0}, is called the likelihood function of the continuous Markov process.

EXAMPLE:

Let us consider the poisson process, where

$$p_{ij}(\tau, t, \lambda) = \frac{[\lambda(t-\tau)]^{j-i}}{(j-i)!} \exp\left[-\lambda(t-\tau)\right],$$

$$c_j(0,\lambda) = \begin{cases} 1 \; for \; j = 0, \\ 0 \; for \; j \neq 0. \end{cases}$$

Here the parameter λ is unknown. Put $t_0=0$. We have n observations at the moments $t_1,t_2,\ldots,t_n$, where $0<t_1<t_2<\ldots<t_n$. We have observed $j_1,j_2,\ldots,j_n$, respectively. Thus the likelihood function is of the form

$$L = \prod_{k=0}^{n-1} \frac{[\lambda(t_{k+1}-t_k)]^{j_{k+1}-j_k}}{(j_{k+1}-j_k)!} \exp[-\lambda(t_{k+1}-t_k)].$$

Hence

$$\log L = -\lambda t_n + j_n \log \lambda + G,$$

where G does not depend on λ. Solving $\frac{\partial \log L}{\partial \lambda_i} = 0$ (i=1,2,…,m), we obtain

$$\lambda^* = \frac{j_n}{t_{n.}}$$

Thus λ^* is the average number of signals per unit of time. It is obvious that the estimate λ^* depends only upon the last observation. It is easy to show that $E(\lambda^*) = \lambda$ and $D^2(\lambda^*) = \lambda/t_n$.

CRAMER RAO INEQUALITY:

STATEMENT:

If t is an unbiased estimator for $\gamma(\theta)$, a function of parameter θ, then

$$var(t) \geq \frac{\left\{\frac{d}{d\theta}\gamma(\theta)\right\}^2}{E\left(\frac{\partial}{\partial\theta}\log L\right)^2} = \frac{[\gamma'\theta]^2}{I(\theta)}$$

PROOF:

Let X be a random variable following the probability density function, $f(x, \theta)$ and let L be the likelihood function of the random sample $(x_1,x_2,\ldots,x_n)$ then

$$L = L(x,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

Since L is the joint probability density function of $x_1,x_2,\ldots,x_n$, $\int L(x,\theta)dx = 1$

where $\int dx = \iint \ldots \int dx_1 dx_2 \ldots dx_n$.

Differentiate with respect to θ and using regularity conditions given above we get,

$$\int \frac{\partial}{\partial \theta} L dx = o \quad => \int \frac{\partial \log L}{\partial \theta} L dx = 0$$

$$E\left(\frac{\partial \log L}{\partial \theta}\right) = 0\text{-------(1)}$$

But E= t($x_1, x_2, \ldots, x_n$) be an unbiased estimates of such that

$$E(t) = \gamma(\theta) \quad => \int t.L dx = \gamma(\theta)\text{---------(2)}$$

Diff with respect to θ, we get,

$$\int t \frac{\partial L}{\partial \theta} dx = \gamma'(\theta)$$

$$\int t\left(\frac{\partial}{\partial \theta} \log L\right) L dx = \gamma'(\theta)$$

$$E\left(t\left(\frac{\partial}{\partial \theta} \log L\right)\right) = \gamma'(\theta)\text{-------(3)}$$

$$Cov\left(t, \frac{\partial}{\partial \theta} \log L\right) = E\left(t\left(\frac{\partial}{\partial \theta} \log L\right)\right) - E(t).E\left(\frac{\partial}{\partial \theta} \log L\right) = \gamma'(\theta)\text{-------(4)}$$

From (1) and (3) we have,

$$\{r(x,y)\}^2 \le 1$$

$$\{cov(x,y)\}^2 \le var\ (X) var(Y)$$

$$\left\{Cov\left(t, \frac{\partial}{\partial \theta} \log L\right)\right\}^2 \le var\ t.var\left(\frac{\partial}{\partial \theta} \log L\right)$$

$$\{\gamma'(\theta)\}^2 \le var(t)[E\left(\frac{\partial}{\partial \theta} \log L\right)^2 - \left\{E\left(\frac{\partial}{\partial \theta} \log L\right)\right\}^2]$$

$$\{\gamma'(\theta)\}^2 \le var(t) E\left\{\left(\frac{\partial}{\partial \theta} \log L\right)^2\right\}$$

$$var(t) \ge \frac{\{\gamma'(\theta)\}^2}{E\left\{\left(\frac{\partial}{\partial \theta} \log L\right)^2\right\}}$$

Which is Cramer Rao inequality.

RAO-BLACKWELL THEOREM:

STATEMENT:

Let X and Y be random variable such that E (Y) = μ and var (Y) = σ^2y>0

Let E (Y/X=x) =ϕ(x)

Then (i) E [ϕ(x)] = μ

(ii) var[ϕ(x)] ≤ var(y)

PROOF:

Let $f_{xy}(x,y)$ be the joint probability density function of the random variables X and Y, $f_1(.)$ and $f_2(.)$ the marginal probability density function of X and Y respectively and h(y/x) be the conditional probability density function of Y for given X=x such that

h(y/x) = {f(x,y)/f(x)}

$$\mathrm{E}\left(\frac{\mathrm{Y}}{\mathrm{X}}=\mathrm{x}\right)=\int_{-\infty}^{\infty} y\,\mathrm{h}\left(\frac{\mathrm{y}}{\mathrm{x}}\right)\mathrm{dy}$$

$$=\int_{-\infty}^{\infty} y\frac{f(x,y)}{f(x)}dy$$

$$=\frac{1}{f_1(x)}\int_{-\infty}^{\infty} yf(x,y)dy$$

$$=\emptyset(x)\text{-------(1)}$$

$$\int_{-\infty}^{\infty} yf(x,y)dy=\emptyset(x)f(x)\text{-------(2)}$$

From (1), we see that the conditional distribution of Y give X=x does not depend on the parameter μ. Hence X is sufficient statistic for μ. Also

$E\{\emptyset(r)\}=E\{E(Y/x)\}$

$$=E(Y)$$

$$=\mu \quad \text{-------(*)}$$

$$var(Y)=E[Y-E(Y)]^2=E(Y-\mu)^2$$

$$=E[Y-\emptyset(x)+\emptyset(x)-\mu]^2$$

$$=E\big(Y-\emptyset(x)\big)^2+E(\emptyset(x)-\mu)^2+2E[\{y-\emptyset(x)\}\{\emptyset(x)-\mu\}]$$

$$E[\{y-\emptyset(x)\}\{\emptyset(x)-\mu\}]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{y-\emptyset(x)\}\{\emptyset(x)-\mu\}f(x,y)dxdy$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{y-\emptyset(x)\}\{\emptyset(x)-\mu\}f_1(x)h(y/x)dxdy$$

$$=\int_{-\infty}^{\infty}\{\emptyset(x)-\mu\}\left[\int_{-\infty}^{\infty}\{y-\emptyset(x)\}h(y/x)dy\right]dx$$

But $\int_{-\infty}^{\infty}\{y-\emptyset(x)\}h(y/x)dy=0$

$$E[\{Y-\emptyset(x)\}\{\emptyset(x)-\mu\}]=0$$

$$var(Y) = E(Y - \emptyset(x))^2 + var(\emptyset(X))$$

$$var(Y) \geq var[\emptyset(x)]$$

$$var[\emptyset(x)] \leq var(Y)$$

Hence the theorem.

MAXIMUM LIKELIHOOD FUNCTION:

$x_1, x_2, \ldots, x_n$ is a random sample of size n with density function $f(x,\theta)$ then the maximum likelihood function is defined as

$L = f(x_1,\theta).f(x_2,\theta)\ldots f(x_n,\theta)$

$\frac{\partial L}{\partial \theta} = 0 \qquad \frac{\partial^2 L}{\partial \theta^2} < 0$

$$\log L = \log f(x_1,\theta) + \log f(x_2,\theta) + \cdots + \log f(x_n,\theta)$$

$$= \sum_{i=1}^{n} \log f(x_i,\theta)$$

Differentiate with respect to θ,

$$\frac{1}{L}\frac{\partial L}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{f(x_i,\theta)} f'(x_i,\theta)$$

$$\frac{\partial L}{\partial \theta} = 0 \quad => \frac{1}{L}\frac{\partial L}{\partial \theta} = 0 \quad => \frac{\partial(\log L)}{\partial \theta} = 0$$

This is called maximum likelihood equation provided $\frac{\partial^2 L}{\partial \theta^2} < 0$.

EXAMPLE:

Find maximum likelihood equation for $f(x,\theta) = \frac{x^{p-1} e^{\frac{-x}{\theta}}}{\theta^p \Gamma p}$

$$L = \prod_{i=1}^{n} f(x_i,\theta)$$

$$= \prod_{i=1}^{n} \frac{{x_i}^{p-1} e^{\frac{-x_i}{\theta}}}{\theta^p \Gamma p}$$

$$= \frac{(x_1 x_2 \ldots x_n)^{p-1} e^{-\sum_{i=1}^{n} \frac{x_i}{\theta}}}{\theta^{np} (\Gamma p)^n}$$

$$\log L = (p-1) \log \prod_{i=1}^{n} x_i - \sum_{i=1}^{n} \frac{x_i}{\theta} - np \log \theta - n \log \Gamma p$$

$$\frac{\partial \log L}{\partial \theta} = \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{np}{\theta} = 0$$

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

$$\hat{\theta} = \frac{1}{np}\sum_{i=1}^{n} x_i = \frac{\bar{x}}{p}$$

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – II - Statistical Inference and Stochastic Process – SMT5208

UNIT II THEORY OF HYPOTHESIS

Theory of Hypothesis; Power function and OC function; Errors; Most Powerful test; Uniformly Most Powerful test; Unbiased test; Neymann-Pearson fundamental Lemma; Problem.

Statistical Hypothesis:

A tentative assumption about the distribution of a random variable made with the purpose of specifying some parameters of the population is known as a statistical hypothesis.

Testing of Hypothesis:

The process of applying certain tests to find out whether pre assigned values of the parameters is acceptable in the light of observation in the sample is known as testing of hypothesis.

Simple Hypothesis:

If the statistical hypothesis specifies the population completely, it is called a simple hypothesis; otherwise it is called a composite hypothesis.

Null Hypothesis:

If there is no significant difference between the sample statistic and the corresponding population parameter or between two samples, then such a hypothesis of no difference is called a null hypothesis and is denoted by H_0.

Alternative Hypothesis:

A hypothesis that is different from the null hypothesis is called an alternative hypothesis and is denoted by H_1.

Test of Hypothesis:

A procedure for deciding whether to accept or reject a null hypothesis H_0 is called the test of hypothesis.

Critical and Acceptance Regions:

The region R of rejection of H_0 is called the critical region and the region A of acceptance of H_0 is called the acceptance region.

Errors in Hypothesis:

When a null hypothesis H_0 is tested against an alternative hypothesis H_1 , one of two types of errors likely to occur.

Type I error:

The error committed in rejecting H_0 , when it is really true is called Type I error.

This is similar to a good product being rejected by the consumer and hence Type I error is also known as **Producer's risk**.

Type II error:

The error committed in accepting H_0, when it is false is called type II error.

As this error is similar to that of accepting a product of inferior quality, it is also known as **Consumer's risk**.

The probabilities of committing the type I error and type II errors are called sizes of the errors and denoted by α and β respectively.

P[rejecting H_0 when it is true] = α

P[accepting H_0 when it is false] = β. Since P[accepting H_0 when it is false] = β,

P[rejecting H_0 when it is false] = 1- β, which is called the **power of the test** H_0 against H_1.

Power of a test = 1- p[type II error]. Therefore, power of test is maximum when the probability of type II error is minimum.

Operating Characteristic Function (OC Function):

The OC function denoted by $K(\theta)$ and is given by $K(\theta)=\dfrac{A^{h(\theta)}-1}{A^{h(\theta)}-B^{h(\theta)}}$, where $A=\dfrac{1-\beta}{\alpha}$ and $B=\dfrac{\beta}{1-\alpha}$, where α and β are the probabilities of type I and type II errors and $h(\theta)\neq 0$ is to be determined for each value of θ, so that $E\left[\left\{\dfrac{f(x,\theta_1)}{f(x,\theta_\circ)}\right\}^{h(\theta)}\right]=1$

Best Test:

The best test is defined as the test which is not only controls α at any desired low level but also minimizes β for a fixed α or minimizes the power of the test $1-\beta$ for a given α. The corresponding critical region is called the **Best Critical Region** (BCR) of size α.

The critical region R is the most powerful (MP) critical region of size α and the corresponding test is called the Most Powerful Test of level α for testing a simple null hypothesis $H_\circ:\theta=\theta_\circ$ against the simple alternative hypothesis $H_1:\theta=\theta_1$ where θ is unknown parameter of the population.

The critical region R is called uniformly most powerful (UMP) critical region of size α and the corresponding test is called the **Uniformly Most Powerful Test** of level α for testing a simple null hypothesis $H_\circ:\theta=\theta_\circ$ against a composite alternative hypothesis $H_1:\theta\neq\theta_\circ$ viz $H_1:\theta=\theta_1\neq\theta_\circ$, where θ is unknown parameter of the population.

Unbiasedness for Hypothesis Testing:

A test function ϕ for which the power function β_ϕ^* satisfies the conditions

(i) $\begin{cases}\beta_\phi^*(\theta)\leq\alpha \ \ for\ \theta\in\Omega_H\,(Null\ hypothesis\ \ parameter\ set)\\ \geq\alpha\ \ for\ \theta\in\Omega_H\,(Alternative\ hypothesis\ \ parameter\ set)\end{cases}$ is said to be unbiased.

Whenever an UMP test exists, it is unbiased, since its power cannot fall below to that of the test $\phi(x)=\alpha$.

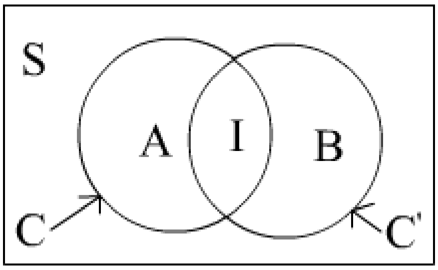
Neymann- Pearson Lemma:

If $\{x_1, x_2,...x_n\}$be a random sample of size n drawn from a population, the distribution of which has the probability density function $f(x,\theta)$so that the joint pdf of $x_1, x_2,...x_n$ is the likelihood function $L(X,\theta)= f(x_1,\theta)f(x_2,\theta)...f(x_n,\theta)$ and if there exists a critical region C of size α and a non-negative constant k such that (i) $\frac{L(X,\theta_1)}{L(X,\theta_\circ)} \geq k,\ for\ all\ X \in C$

(ii) $\frac{L(X,\theta_1)}{L(X,\theta_\circ)} \leq k,\ for\ all\ X \notin C$, then C is the best critical region of size α for testing a simple null hypothesis $H_\circ : \theta = \theta_\circ$ against a simple alternative hypothesis $H_\circ : \theta = \theta_1$.

Proof:

Let C' be any other critical region of size $\leq \alpha$ .The critical regions C and C' have intersection region I and non-intersecting parts are denoted by A and B respectively.



Since size of the critical region C =α (given), p[Reject $H_\circ / H_\circ$ is true]= α

p[$X \in C / H_\circ$ is true]= α

$$\int_C L(X,\theta_\circ)dX = \alpha$$

Similarly, $\int_{C'} L(X,\theta_\circ)dX \leq \alpha$................(1)

By step (1), $\int_C L(X,\theta_\circ)dX \geq \int_{C'} L(X,\theta_\circ)dX$

$$\int_A L(X,\theta_\circ)dX + \int_I L(X,\theta_\circ)dX \geq \int_I L(X,\theta_\circ)dX + \int_B L(X,\theta_\circ)dX$$

Cancelling the common integrals we get,

$\int_A L(X,\theta_\circ)dX \geq \int_B L(X,\theta_\circ)dX$.................(2)

Let β be the size of the type II error for C.

P[accept $H_\circ / H_1$ is true]= β. Therefore, p[Reject $H_\circ / H_1$ is true] = $1-\beta$

p[$X \in C / H_1$ is true]= $1-\beta$

$$\int_C L(X,\theta_1)dX = 1-\beta \text{....................(3)}$$

Similarly, $p\{X \in C^{'} / H_1 is\ true\} = 1-\beta^{'}$

Where $\beta^{'}$ is the size of the type II error for $C^{'}$

$$\int_{C^{'}} L(X,\theta_1)dX = 1-\beta^{'} \text{....................(4)}$$

(3) – (4) gives

$$\beta^{'} - \beta = \int_C L(X,\theta_1)dX - \int_{C^{'}} L(X,\theta_1)dX$$

$$\beta^{'} - \beta = \int_A L(X,\theta_1)dX + \int_I L(X,\theta_1)dX - \int_I L(X,\theta_1)dX - \int_B L(X,\theta_1)dX$$

$$\beta^{'} - \beta = \int_A L(X,\theta_1)dX - \int_B L(X,\theta_1)dX \text{.....................(5)}$$

From the given condition (i)

$$L(X,\theta_1) \geq kL(X,\theta_\circ),\ for\ all\ X \in C$$

Hence $\int_A L(X,\theta_1)dX \geq k\int_A L(X,\theta_\circ)dX$.....................(6)

From the given condition (ii)

$$L(X,\theta_1) \leq kL(X,\theta_\circ),\ for\ all\ X \notin C$$

Hence $\int_B L(X,\theta_1)dX \leq \int_B L(X,\theta_\circ)dX$

(or) $-\int_B L(X,\theta_1)dX \geq -k\int_B L(X,\theta_\circ)dX$.................(7)

Using (6) and (7) in (5), we get

$$\beta^{'} - \beta \geq k\int_A L(X,\theta_\circ)dX - k\int_B L(X,\theta_\circ)dX$$

$\beta^{'} - \beta \geq 0$ by step (2) or $1-\beta > 1-\beta^{'}$

(i.e) C has greater power than C’ and hence C is the best critical region.

Hence the proof.

Problems:

1. Test the hypothesis $H_\circ : \theta = 1.5$ *against* $H_1 : \theta = 2.5$ by using a single observation X, given that the density function of X is given by $f(x,\theta) = \frac{1}{\theta}; 0 \le x \le \theta \text{ and } = 0;$ otherwise, if $x \ge 0.75$ is taken as the critical region, then obtain the sizes of Type I and Type II errors and the power function of the test.

Solution:

α = size of type I error.

$= p\{H_\circ \text{is rejected} / H_\circ \text{is true}\}$

$= p\{X \in W / H_\circ \text{is true}\}$

$= p\{X \ge 0.75 / \theta = 1.5\}$, since $W = p\{X \ge 0.75\}$

$$= p\{0.75 \le X \le 1.5\} = \int_{0.75}^{1.5} \frac{1}{1.5} dX = \frac{2}{3}(1.5 - 0.75) = 0.5$$

β = size of typeII error

$= p\{H_\circ \text{is accepted} / H_1 \text{is true}\}$

$= p\{X \notin W / H_1 \text{is true}\}$

$$= p\{0 \le X \le 0.75 / \theta = 2.5\} = \int_{0}^{0.75} \frac{1}{2.5} dX = \frac{1}{2.5}(0.75) = 0.3$$

Power function of the test = $1 - \beta$ =1-0.3 = 0.7

2. If a random variable X has the density function $f(x,\theta) = \theta x^{\theta - 1} : 0 < x < 1$ and = 0, elsewhere, test the simple hypothesis $H_\circ : \theta = 1$ against the alternative simple hypothesis $H_1 : \theta = 2$, using a simple random sample (x_1, x_2) of size 2, taking the critical region as $W = \{(x_1, x_2) : x_1 + x_2 \le 1\}$. Find α and the power function of the test.

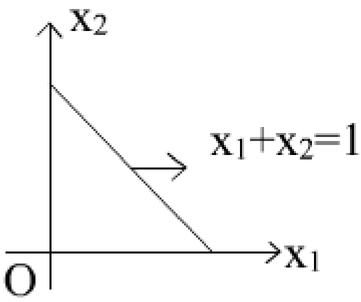
   Solution:

   Give X be a random variable with density function $f(x,\theta) = \{\theta x^{\theta - 1} : 0 < x < 1$, = 0 elsewhere.

   Under $H_\circ : \theta = 1$, $f(x_1, x_2, \theta) = 1, 0 < x_1, x_2 < 1$

   Under $H_1 : \theta = 2$, $f(x_1, x_2, \theta) = 2x_1^{2-1} x_2^{2-1} = 4x_1 x_2$, $0 < x_1, x_2 < 1$

   $\alpha = p\{X \in W / H_\circ \text{is true}\}$



$$\alpha = p\{x_1 + x_2 \le 1/\theta = 1\} = \iint_W f(x_1, x_2, 1)dx_1 dx_2$$

$$= \int_0^1 \int_0^{1-x_2} (1) dx_1 dx_2 = \text{Area of the triangle OAB} = ½$$

$$1-\beta = p\{\text{Re } jecting H_\circ / H_1 is\ true\}$$

$$= p\{X \in W / H_1 is\ true\}$$

$$= p\{x_1 + x_2 \le 1/\theta = 2\}$$

$$= \iint_W f(x_1, x_2, \theta = 2)dx_1 dx_2 = \int_0^1 \int_0^{1-x_2} 4x_1 x_2 dx_1 dx_2 = 4\int_0^1 \left(\frac{x_1^2}{2}\right)_0^{1-x_2} x_2 dx_2$$

$$= 2\int_0^1 (1-x_2)^2 x_2 dx_2 = 2\int_0^1 (1 - 2x_2 + x_2^2) x_2 dx_2 = 2\int_0^1 (x_2 - 2x_2^2 + x_2^3) dx_2$$

$$= 2\left[\frac{x_2^2}{2} - \frac{2x_2^3}{3} + \frac{x_2^4}{4}\right]_0^1 = 2\left[\frac{1}{2} - \frac{2}{3} + \frac{1}{4}\right]$$

$$\therefore \quad 1-\beta = \frac{1}{6}$$

Likelihood Ratio Test (LRT)

The LRT which can be considered as a generalization of Neymann Pearson lemma is a test used for testing a simple or composite hypothesis $H_\circ$ against a simple or composite hypothesis H_1.

If $L(X,\theta)$ is the likelihood function for a random sample $X = \{x_1, x_2, \ldots x_n\}$ drawn from the population with density function $f(x,\theta); \theta \in \Theta$, the likelihood ratio statistic $\lambda(X)$ defined by $\lambda(X) = \dfrac{\underset{\theta \in \Theta_\circ}{Max}\{L(X,\theta)\}}{\underset{\theta \in \Theta_1}{Max}\{L(X,\theta)\}}$ is used for testing $H_\circ$ against H_1.

The Likelihood Ratio Test is then stated as, " Reject $H_\circ$ if $\lambda(X) < k$, where k is usually determined by the inequality that $p\{Type\ I\ error\} \le \alpha$.

Problem:

1. Find the likelihood ratio test of $H_\circ : \theta = \theta_\circ$ against $H_1 : \theta \neq \theta_\circ$ based on a sample of size 1, drawn from the population with density $f(x,\theta) = \dfrac{2(\theta - x)}{\theta^2}, 0 < x < \theta$

Solution:

$$L(X,\theta) = f(x,\theta) = \frac{2(\theta - x)}{\theta^2}$$

The Maximum Likelihood Estimatoe (MLE) is given by $\dfrac{\partial}{\partial\theta}(\log L) = 0$

$$\Rightarrow \frac{\partial}{\partial\theta}[\log 2 + \log(\theta - x) - 2\log\theta] = 0$$

$$\Rightarrow \frac{1}{(\theta - x)} - \frac{2}{\theta} = 0$$

$$\Rightarrow \theta - 2(\theta - x) = 0$$

$$\Rightarrow -\theta - 2x = 0 \Rightarrow \theta = 2x$$

$$\hat{\theta} = 2x$$

The likelihood ratio statistic $\lambda(X)$ is given by $\lambda(X) = \dfrac{MaxL(X,\theta_\circ)}{MaxL\left(X,\hat{\theta}\right)}$..............(1), since

$H_\circ : \theta = \theta_\circ$.

Maximum of $L(X,\theta) = f(x,\theta) = \dfrac{2(\theta - x)}{\theta^2}$ is given by $\dfrac{\partial L}{\partial\theta} = 0$

$$\Rightarrow \frac{2[\theta^2(1) - (\theta - x)2\theta]}{-\theta^4} = 0$$

$$\Rightarrow \theta - 2(\theta - x) = 0$$

$$\Rightarrow -\theta - 2x = 0 \Rightarrow \theta = 2x$$

Therefore, $MaxL\left(X,\hat{\theta}\right) = \dfrac{2(2x - x)}{4x^2} = \dfrac{1}{2x}$..............(2)

Using (2) in (1) we get

$$\lambda(X) = \frac{4x(\theta_\circ - x)}{\theta_\circ^{\,2}}$$

$\therefore$ As per the likehood ratio test, $H_\circ$ is rejected, if $\lambda(X) < k$

$$x(\theta_\circ - x) < k\frac{\theta_\circ^{\,2}}{4} = b,\ say$$

$$\Rightarrow -x^2 + \theta_\circ x - b < 0 \Rightarrow x^2 - \theta_\circ x + b > 0$$

$$\left(x - \frac{\theta_\circ}{2}\right)^2 > \frac{\theta_\circ^{\,2}}{4} - b\left(= a^2, say\right)$$

$\left(x - \frac{\theta_\circ}{2}\right) > a$, which gives the critical region associated with the test, a is found out by the size α of the critical region.

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

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – III - Statistical Inference and Stochastic Process – SMT5208

UNIT III -NON PARAMETRIC TESTS

Non-Parametric test; Introduction; Kolomogorov Smirnov test for two samples; sign test; Wald –Wolfowiz Run test; Median test for two samples and Mann-Whitney U-test.

NON PARAMETRIC TEST:

The test that do not make restrictive assumptions about the shape of the population distributions are known as non parametric test. The hypothesis of a non parametric test is concerned with something other than the value of a population parameter.

KOLMOGOROV-SMIRNOV TEST:

Suppose that a company has conducted a field survey covering 200 respondents. Apart from other questions it asked the respondent to indicate on a 5 point scale how much the durability of a particular product is important to them.

The respondent indicated as follows:

Very important : 50

Somewhat important : 60

Neither important nor unimportant: 20

Somewhat unimportant : 40

Very unimportant : 30

Use the Kolmogorov test to test the hypothesis that there is no difference in importance rating for durability among the respondent.

SOLUTION:

Let H_0: There is no difference in importance rating for durability among the respondents.

H_1: There is a difference in importance rating for durability among the respondents.

OBSERVED NUMBERS	**OBSERVED PROPORTION**	**OBSERVED CUMMULATIVE PROPORTION (F_o)**	**NULL (OR) EXPECTED PROPORTION**	**CUMMULATIVE NULL PROPORTION (F_e)**	**D_n=Max $\lvert F_e - F_o \rvert$**
50	50/200=0.25	0.25	0.2	0.2	0.05
60	60/200=0.30	0.55	0.2	0.4	0.15
20	20/200=0.10	0.65	0.2	0.6	0.05
40	40/200=0.20	0.85	0.2	0.8	0.05

30	30/200=0.15	1.0	0.2	1.0	0
200					

D_n= Max $|F_e - F_o|$

= 0.15

Calculated D_n= 0.15

Since N= 200>30

The tabulated value of $D_n = \frac{1.36}{\sqrt{N}} = \frac{1.36}{\sqrt{200}}$ = 0.0962

Since the calculated value of D_n greater than the tabulated value of D_n. Hence H_o is rejected.

Therefore there is significant difference in importance rating for durability among the respondents.

SIGN TEST:

Consider the following data:

SAMPLE1	24.1	22.9	23.0	26.1	25.0	30.8	27.1	23.2	22.8
SAMPLE 2	23.7	24.6	30.3	23.9	21.8	28.1	25.4	31.2	30.9

Test whether 30 can be regarded as population median.

SOLUTION:

Let H_0: $M=M_0$

H_1: $M \neq M_0$

Since M_0=30,

-,-,-,-,-,+,-,-,-,-,-,+,-,-,-,-,+,+.

S_+= No. of + signs=4

S_-= No. of – signs=14

Calculated value x= min(S_+,S_-)

= min(4,14)

= 4

By using Binomial table B(n,x,p)= B(n=18,x=4,p=0.5)

Tabulated Value= p= 0.0154

Calculated Value > Tabulated Value

H_0 is accepted.

ie) $M = M_0$

Therefore, 30 can be regarded as the population.

WALD WOLFOWIZ RUN TEST:

The personal director of a company wishes to select applicants for advanced training without regard to sex. Let W denote women and M denote men and the pattern of arrival be

M WWW MMM WW M WWWW MMMM W M W MM WWW MM W MMMM WW M WW MMMM WW M WWWW MM WW M W M WW

Will you conclude that the applicants have arrived in random fashion.

SOLUTION:

Let H_0: The applicants have arrived in random fashion.

H_1: The applicants have not arrived in random fashion.

The test statistic is $z = \frac{r - \mu_r}{\sigma_r}$

Where, $\mu_r = \frac{2n_1 n_2}{n_1 + n_2} + 1$

$$\sigma_r = \sqrt{\frac{2n_1 n_2(2n_1 n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

and r= no. of runs= 28

n_1= no. of observations I= 28

n_2 = no. of observations II= 30

$$\mu_r = \frac{2(28)(30)}{28+30} + 1$$

$$\mu_r = 29.97$$

$$\sigma_r = \sqrt{\frac{2(28)(30)(2(28)(30)-28-30)}{(28+30)^2(28+30-1)}}$$

$$\sigma_r = 3.7698$$

$$z = \frac{28-29.97}{3.7698}$$

z_{cal}=-0.5226

$|z_{cal}|$ = 0.5226

Tabulated value of z at 5% LOS = 1.96

Z_{tab}= 1.96

Since $z_{cal} < Z_{tab}$

H_0 is accepted.

Therefore applicants have arrived in random fashion.

MEDIAN TEST:

Do urban and rural junior high school students differ with respect to their level of mental help

URBAN	**35 26 27 21 27 38 23 25 25 27 45 46 33 26 46 41**
RURAL	**29 50 43 22 42 47 42 32 50 37 37 34 31**

SOLUTION:

Let H_0: Urban and rural high school students do not differ in their mental help ie)$M_U = M_R$

H_1: $M_U \neq M_R$.

Arrange the given data of the samples in ascending order.

21,22,23,25,25,26,26,27,27,27,29,31,32,33,34,35,37,37,38,41,42,42,43,45,46,46,47,50,50.

Median $= \left(\frac{n+1}{2}\right)^{th}$ item [since n is odd]

$= \left(\frac{29+1}{2}\right)^{th}$ item

$= 15^{th}$ item

Median = 34

	URBAN	**RURAL**	**TOTAL**
Urban and rural values above median	6	8	14
Urban and rural values below median	10	4	14
TOTAL	16	12	28

The test statistics is,

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi^2 = \frac{28\,(24-80)^2}{(14)(14)(16)(12)} = 2.33$$

The tabulated value of χ^2 at 5% LOS with (R-1)(C-1) = 1 degrees of freedom is $\chi^2_{tab} = 3.841.$

$$\chi^2_{cal} < \chi^2_{tab}$$

H_0 is accepted.

$M_U = M_R$

Rural and urban high school students do not differ in their mental help.

MANN-WHITNEY U-TEST:

From the following table:

SAMPLE I	**70 90 82 64 86 77 84 79 82 89 73 81 83 66**
SAMPLE II	**86 78 90 82 65 87 80 88 95 85 96 76 94**

Test whether the two samples have significant difference in effectiveness. Use Mann-Whitney U-Test at 5% Los.

SOLUTION:

Let H_0 : There is no significant difference between two samples in effectiveness.

H_1 : There is no significant difference between two samples in effectiveness.

The test statistic is $z = \frac{U-\mu}{\sigma}$

SAMPLE I	RANK OF I	SAMPLE II	RANK OF II
70	4	86	18.5
90	23.5	78	8
82	13	90	23.5
64	1	82	13
86	18.5	65	2
77	7	87	20
84	16	80	10
79	9	88	21
82	13	95	26
89	22	85	17
73	5	76	6
81	11	94	25
83	15		

66	3		
	$R_1 = 161$		$R_2 = 190$

Given $n_1 = 14$, $n_2 = 12$

Since $R_1 < R_2$

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$= (14)(12) + \frac{14(14+1)}{2} - 161$$

$U = 112$

$$\mu = \frac{n_1 n_2}{2} = \frac{14 \times 12}{2} = 84$$

$$\sigma = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}} = \sqrt{\frac{14 \times 12(14 + 12 + 1)}{12}} = 19.44$$

$$z = \frac{U - \mu}{\sigma} = \frac{112 - 84}{19.44}$$

$$z_{cal} = 1.44$$

The tabulated value of z at 5% LOS=1.96

Since $z_{cal} < z_{tab}$. H_0 is accepted.

Therefore there is no significant difference between two samples in effectiveness.

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – IV - Statistical Inference and Stochastic Process – SMT5208

UNIT IV ANALYSIS OF VARIANCE

Analysis of variance, One-way classification; Two way classification; Principles of Experimental design; CRD, RBD and LSD and simple problems.

DESIGN OF EXPERIMENTS AND ANALYSIS OF VARIANCE:

The main aim of the design of experiments is to control the contribution of extraneous variables (fertility of the soil, quality of the seed and hence to minimize the experimental error so that the results of the experiments used and the amount of rain fall) could be attributed only to the experimental variables. [Quantity of the manure used and the amount of yield of paddy].

BASIC PRINCIPLE'S OF EXPERIMENTAL DESIGNS:

The following three principles are adopted while designing experiments are:

- Randomization
- Replication
- Local control

RANDOMISATION:

Plots

- Experimental Group
- Control Group

In any information regarding extraneous variables and the nature and magnitude of their effect on the response variable in question is not available, we resort to randomisation which means selection of plots for the experimental and control group in a random manner. This technique provides the most effective way of eliminating any unknown bias in the experiments.

REPLICATION:

In order to estimate the amount of experimental error and hence to get some idea of the precision of the estimates of the manure effects it is essential to carry out more than one test on each manure.

LOCAL CONTROL:

In order to achieve adequate control of extraneous variable, another important principle used in the experimental design is the local control. This includes techniques such as grouping, blocking and balancing of the experimental units (plots) used in the experimental design.

GROUPING:

Grouping means combining sets of homogeneous plots into groups so that different manures may be used in the different groups need necessarily be the same.

BLOCKING:

Assigning the same number of plots in different blocks. The plots in the same block may be assumed to relatively to be relatively homogeneous. We use as many types of manure as the number of plots in a block in a random manner.

BALANCING:

Balancing means making minor change in the procedure of grouping and blocking and then assigning the manures in such a manner that a balanced configuration is obtained.

SOME BASIC DESIGN OF EXPERIMENTS:

- Completely Randomised Design (CRD) – One Way Classification.
- Randomised Block Design (RBD) – Two Way Classification.
- Latin Square Design (LSD) – Three Way Classification.

COMPLETELY RANDOMISED DESIGN (CRD):

CRD is a design in which N values of a given random variable 'X' (which may represent, for example the yield of paddy) contained in a sample are subdivided into the changes according to one factor of classification (say different manures).

RANDOMISED BLOCK DESIGN (RBD):

RBD is a design in which the N variable value (yield of paddy) is classification according to two factors.

LATIN SQUARE DESIGN (LSD):

LSD is a design in which $N=n^2$ plots are taken and arranged in the form of an $n \times n$ square such that the plots in each row will be homogeneous as far as possible with respect to one factor of classification say soil fertility. Plots in each column will be homogeneous as far as possible with respect to another factor of classification say seed quality. Then n treatment (Third factor of classification) represented by letters are given to these plots such that each treatment occurs only one in each row and only once in each column. The various possible arrangements obtained in this manner are known as Latin Square of the Order n.

ANALYSIS OF VARIANCE (ANOVA):

After planning and conducting experiments, the results obtained must be analysed and interpreted. The technique for making statistical inference is known as the analysis of variance.

PROBLEMS:

ONE WAY ANOVA

There are 3 main brands of a certain powder. The set of 110 sample value is examine and found to be allocated among 4 groups (A, B, C, D) and these brands (I, II, III) as shown below:

BRANDS	A	B	C	D

	GROUPS			
I	0	4	8	15
II	5	8	13	6
III	8	19	11	13

Is there any significant difference in Brand's preference?

SOLUTION:

Let H_0: There is no significant difference in brands preference.

H_1: There is significant difference in brands preference.

STEP:1

N= Total number of observations =12

T= Sum of all observations = 110

Correction factor CF=$\frac{T^2}{N}$

$=\frac{(110)^2}{12}$

CF=1008.33

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2
0	5	8	0	25	64
4	8	19	16	64	361
8	13	11	64	169	121
15	6	13	225	36	169
$\sum x_1$=27	$\sum x_2$=32	$\sum x_3$=51	$\sum x_1^2$=305	$\sum x_2^2$=294	$\sum x_3^2$=715

STEP 2:

SST = $(\sum x_1^2 + \sum x_2^2 + \sum x_3^2)$-CF

= (305+294+715)-1008.33

SST = 305.67

SSC = Sum of squares between columns

= $(\frac{(\sum x_1^2)^2}{n_1}+\frac{(\sum x_2^2)^2}{n_2}+\frac{(\sum x_3^2)^2}{n_3})$-CF

SSC = 80.17

STEP 3:

SSE = Error sum of squares

= SST-SSC

=225.5

ANOVA TABLE:

SOURCE OF VARIANCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SUM OF SQUARES	VARIANCE RATIO
BETWEEN THE SAMPLE COLUMN	SSC = 80.17	C-1 = 2	MSC = 40.085	F_C = 1.60 $\therefore MSC > MSE$
WITHIN THE SAMPLE	SSE = 225.5	N-C = 9	MSE = 25.06	

Tabulated value of F at 5% LOS with (γ_1 , γ_2)

Since $(\gamma_1 > \gamma_2)$ degrees of freedom is $F_{tabulated}$ = 4.26

Since $F_{calculated} < F_{tabulated}$

$\therefore$ H_0 is accepted

$\therefore$ There is no significant difference in brand preference.

TWO WAY ANOVA

Three types of indoor lightings A_1, A_2, A_3 where sprayed on 3 types of flowers F_1, F_2, F_3 grown from seed. The height in (on after 12 weeks of growth are given below. Analyze the data

	FLOWERS			
		F_1	F_2	F_3
LIGHTING	A_1	16	24	19
	A_2	15	25	18
	A_3	21	31	15

SOLUTION:

H_0: There is no significant different in growth of the flowers. There is no significance difference in brightness of lightings.

H_1:There is significant different in growth of the flowers. There is no significance difference in brightness of lightings.

STEP I:

N= Total number of observations = 9

T= Sum of all observations = 184

Correction factor CF=$\frac{T^2}{N}$

$=\frac{(184)^2}{9}$

CF= 3761.78

Y/X	X_1	X_2	X_3		X_1^2	X_2^2	X_3^2	
Y_1	16	24	19	59	256	576	361	1193
Y_2	15	25	18	58	225	625	324	1174
Y_3	21	31	15	67	441	961	225	1627
	$\sum x_1$= 52	$\sum x_2$=80	$\sum x_3$=52		$\sum x_1^2$= 922	$\sum x_2^2$=2162	$\sum x_3^2$=910	

STEP II:

SST = $(\sum x_1^2 + \sum x_2^2 + \sum x_3^2)$-CF

SST = 232.22

SSC = Sum of squares between columns

= $(\frac{(\sum x_1^2)^2}{n_1}+\frac{(\sum x_2^2)^2}{n_2}+\frac{(\sum x_3^2)^2}{n_3})$-CF

SSC = 174.21

SSR= Sum of squares between rows

$SSR = (\frac{(\sum Y_1^2)^2}{n_1}+\frac{(\sum Y_2^2)^2}{n_2}+\frac{(\sum Y_3^2)^2}{n_3})$-CF

SSR = 16.22

STEP III:

SSE = Error sum of squares

= SST-SSC-SSR

=41.79

ANOVA TABLE:

SOURCE OF VARIANCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SUM OF SQUARES	VARIANCE RATIO
BETWEEN THE SAMPLE COLUMN	SSC = 174.21	C-1 = 2	MSC = 87.105	$F_C = 8.34$ $\therefore MSC > MSE$
BETWEEN THE SAMPLE OF ROWS	SSR = 16.22	R-1 = 2	MSR = 8.11	
WITHIN SAMPLE	SSE = 41.79	(C-1) (R-1) = 4	MSE = 10.44	$F_R = 0.77$

Tabulated value of F at 5% LOS with (2, 4) degree of freedom= 6.94 (between columns)

Since $F_{cal(c)} > F_{tab}$

Therefore, H_0 is rejected.

Tabulated value of F at 5% LOS with (4, 2) degree of freedom= 19.2 (between rows)

Since $F_{cal(R)} < F_{tab}$

Therefore, H_0 is accepted.

THREE WAY ANOVA

The following is a Latin square of a design where 4 vertices of seeds are being tested setup the analysis of variance table and state your conclusion show that you may carryout suitable change of origin and table.

A105	B95	C125	D115
C115	D125	Q105	B105
D115	C95	B105	A115
B95	A135	B95	C115

SOLUTION:

Take origin = 100 and scale = 5

Let H_0: There is no significant difference between rows, columns and treatments.

H_1: There is significant difference between rows, columns and treatments.

STEP I:

N= Total number of observations = 16

T= Sum of all observations = 32

Correction factor CF=$\frac{T^2}{N}$

$=\frac{(32)^2}{16}$

CF = 64

$Y/_X$	X_1	X_2	X_3	X_4		X_1^2	X_2^2	X_3^2	X_4^2	
Y_1	1	-1	5	3	8	1	1	25	9	36
Y_2	3	5	1	1	10	9	25	1	1	36
Y_3	3	-1	1	3	6	9	1	1	9	20
Y_4	-1	7	-1	3	8	1	49	1	9	60
	6	10	6	10	32	20	76	28	28	

BETWEEN TREATMENTS:

A	B	C	D
1	-1	5	3
1	1	3	5
3	1	-1	3
7	-1	3	-1
12	0	10	10

STEP II:

SST = Total number of squares

SST = $(\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2)$-CF

SST = 88

SSC = Sum of squares between columns

$= (\frac{(\sum x_1^2)^2}{n_1}+\frac{(\sum x_2^2)^2}{n_2} + \frac{(\sum x_3^2)^2}{n_3} + \frac{(\sum x_4^2)^2}{n_4})$-CF

SSC = 4

SSR= Sum of squares between rows

$SSR = (\frac{(\sum Y_1^2)^2}{n_1}+\frac{(\sum Y_2^2)^2}{n_2} + \frac{(\sum Y_3^2)^2}{n_3} + \frac{(\sum Y_4^2)^2}{n_4})$-CF

SSR = 2

SSk= Sum of squares between rows

$$SSk = \left(\frac{(\sum z_1^2)^2}{n_1} + \frac{(\sum z_2^2)^2}{n_2} + \frac{(\sum z_3^2)^2}{n_3} + \frac{(\sum z_4^2)^2}{n_4}\right) - CF$$

SSK = 22

STEP III:

SSE = Error sum of squares

= SST-SSC-SSR-SSK

=60

ANOVA TABLE:

SOURCE OF VARIANCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SUM OF SQUARES	VARIANCE RATIO
BETWEEN THE SAMPLE COLUMN	SSC = 4	C-1 = 3	MSC = 1.33	F_C = 7.52 $\therefore MSC > MSE$
BETWEEN THE SAMPLE OF ROWS	SSR = 2	R-1 = 3	MSR = 0.67	
BETWEEN THE SAMPLE OF TREATMENTS	SSK = 22	K-1 = 3	MSK = 7.33	F_R = 14.93
WITHIN SAMPLE	SSE = 60	(K-1) (K-2) = 6	MSE = 10	F_K=1.36

Tabulated value of F at 5% LOS with (6, 3) degree of freedom= 4.76 (between columns)

Since $F_{cal(c)} > F_{tab}$

Therefore, H_0 is rejected.

Tabulated value of F at 5% LOS with (6, 3) degree of freedom= 4.76 (between rows)

Since $F_{cal(R)} > F_{tab}$

Therefore, H_0 is rejected.

Tabulated value of F at 5% LOS with (6, 3) degree of freedom= 4.76 (between treatments)

Since $F_{cal(K)} < F_{tab}$

Therefore, H_0 is accepted.

Therefore, we conclude that, there is no significant difference in treatments and there is significant difference in columns and rows.

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – V - Statistical Inference and Stochastic Process – SMT5208

UNIT V STOCHASTIC PROCESS

Stochastic processes; specification of SP; Markov chain; transition probabilities; Determination of higher order transition probabilities; Chapman-Kolmogorov equation; Poisson processes (introduction only).

RANDOM PROCESSES:

A random process is a collection of random variables {X(s, t)} that are functions of real variable‘t’ where s ϵ S, S is the sample space and t ϵ T, T is an index set.

CLASSIFICATION OF RANDOM PROCESS:

- Discrete random sequence
- Continuous random sequence
- Discrete random process
- Continuous random process.

MARKOV PROCESS:

An important class of stochastic process is a Markov process which satisfies the Markovian property that the future value depends on the present value and not on the past values is called a Markov process.

MARKOV CHAIN:

If for all n,

$P(X_n=a_n/X_{n-1}=a_{n-1}, X_{n-2}=a_{n-2},\dots X_0=a_0) = P(X_n=a_n/X_{n-1}=a_{n-1})$

Then the process (X_n= 0,1,2,3…) is called a Markov chain.

STATES OF MARKOV CHAIN:

$a_1, a_2,\dots,a_n$ are called the states of the Markov chain.

ONE STEP TRANSITION PROBABILITY:

The one step transition probability is denoted by $p_{ij}(n-1,n)$ and is defined by the conditional probability from state a_i to state a_j at the nth step $p\{X_n = a_j / X_{n-1} = a_i\}$.

HOMOGENEOUS MARKOV CHAIN:

If the one- step transition probability does not depend on the step, (i.e.) $p_{ij}(n-1,n) = p_{ij}'(m-1,m)$, Then the Markov chain is called a homogeneous Markov chain or the chain is said to have stationary transition probabilities.

TRANSITION PROBABILITY MATRIX (TPM) :

When the Markov chain is homogeneous, the one-step transition probability is denoted by p_{ij}. The matrix $P = \{p_{ij}\}$ is called one step transition probability matrix. The transition probability matrix is a stochastic matrix with row sum is equal to 1.

n-STEP TRANSITION PROBABILITY:

The conditional probability $P_{ij}^{(n)} = P(X_n=a_j/X_0=a_i)$ is called the n-step transition probability.

CLASSIFICATION OF STATES OF MARKOV CHAIN:

IRREDUCIBLE CHAIN:

If $p_{ij}^{(n)} > 0$ for some n and for all i and j , then every state can be reached from every other state. When this condition is satisfied, the Markov chain is said to be irreducible.

RETURN STATE:

State i of Markov chain is called a return state, if $p_{ii}^{(n)} > 0$ for some $n > 1$.

PERIODIC STATE:

The period d_i of a return state i is defined as the greatest common divisor of all m such that $p_{ii}^{(m)} > 0$. State i is said to be periodic with period $d_i > 1$. State i is said to be aperiodic if $d_i = 1$.

FIRST RETURN TIME PROBABILITY:

The probability that the chain returns to state i , having started from state i for the first time at the nth step is denoted by $f_{ii}^{(n)}$ and called the first return time probability or the recurrence time probability.

RECURRENT STATE:

A state i is said to be persistent or recurrent if the return to state i is certain (i.e.) if $F_{ii} = 1$. The state i is said to be transient if the return to state i is uncertain (i.e.) if $F_{ii} < 1$.

NON-NULL PERSISTENT STATE:

The state i is said to be non-null persistent if its mean recurrence time μ_{ii} is finite and null persistent if $\mu_{ii} = \infty$.

ERGODIC STATE:

The state i is said to be ergodic if it is non-null persistent and aperiodic.

Chapman-Kolmogorov Theorem:

If P is the tpm of a homogeneous Markov chain, then the n-step tpm $P^{(n)}$ is equal to P^n. (i.e) $p_{ij}^{(n)} = \left[p_{ij}\right]^n$.

POISSON PROCESS:

If {X(t)} represent the no. of occurrences of certain event in (o,t). Then , the discrete random process {X(t)} is called the poisson process, provided the following postulates are satisfied,

(i) P[1 occurrence in (t, t+h)] = λh+o(h)

(ii) P[0 occurrence in (t, t+h)] = 1-λh+o(h)

(iii) P[more than 1 occurrence in (t, t+h)] = o(h)

(iv) $\lim_{n \to 0} \frac{o(h)}{h} = 0$

(v) Events occur in non-overlapping time intervals.

PROBLEMS:

1. A man either drives a car or catches a train to go to office each day. He never goes two days in row by train. But iff he drives one day, then the next day he is just as likely to drive again as he is to travel by train. Now suppose, that on the first day of the week, the man tossed a fair dice and drove to work. Iff a six appeared. Find (i) The probability that he takes a train on the third day. (ii) The probability that he drives to work in the long run.

SOLUTION:

The states space of a Markov chain is {Train, Car} i.e. S = {T, C}

The transition probability matrix corresponding to the chain is given by

$$P = \begin{pmatrix} o & 1 \\ 1/2 & 1/2 \end{pmatrix}$$

P [driving to work] = 1/6

P [travel by train] = 1- 1/6 = 5/6

The initial probability is

$P^{(1)}$ = (5/6,1/6)

$P^{(2)} = P^{(1)} P = (5/6,1/6)\begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$

= (1/12,11/12)

$P^{(3)} = P^{(2)} P = (1/12,11/12)\begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$

= (11/24,13/24)

(i) P[He takes the train on the third day]= 11/24

(ii) Probability in the long run is

$$\pi P = \pi$$

$$(\pi_1, \pi_2)\begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix} = (\pi_1, \pi_2)$$

$$\left(\frac{\pi_2}{2}, \pi_1 + \frac{\pi_2}{2}\right) = (\pi_1, \pi_2)$$

$$\frac{\pi_2}{2} = \pi_1 \text{ ---------- (1)}$$

$$\pi_1 + \frac{\pi_2}{2} = \pi_2 \text{ ----------- (2)}$$

and $\pi_1 + \pi_2 = 1$--------- (3)

Solving equations (1),(2) and (3) we get,

$$\frac{\pi_2}{2} + \pi_2 = 1$$

$$\frac{\pi_2 + 2\pi_2}{2} = 1$$

$$3\pi_2 = 2$$

$$\pi_2 = 2/3$$

Sub π_2 in eqn (3) we get,

$$\pi_1 + 2/3 = 1$$

$$\pi_1 = 1/3$$

$$\pi = (1/3, 2/3)$$

P(He drives to work in the long run) = 2/3.

2. A housewife buys 3 kinds of cereals A, B and C. She never buys the same cereal in successive weeks. If she buys cereal A, the next week she buys cereal B. However if she buys B or C, the next week she is 3 times as likely to buy A as the other cereal in the long time, how often does she buy each of the 3 cereals in the long time.

SOLUTION:

The state space of a Markov chain is S={A,B,C}

The transition probability matrix is given by

$$P=\begin{pmatrix} 0 & 1 & 0 \\ 3/4 & 0 & 1/4 \\ 3/4 & 1/4 & 0 \end{pmatrix}$$

In the long run, $\pi P = \pi$

$$(\pi_1\pi_2\pi_3)\begin{pmatrix} 0 & 1 & 0 \\ 3/4 & 0 & 1/4 \\ 3/4 & 1/4 & 0 \end{pmatrix} = (\pi_1\pi_2\pi_3)$$

$\frac{3\pi_2}{4}+\frac{3\pi_3}{4}=\pi_1$---------(1)

$\pi_1+\frac{\pi_3}{4}=\pi_2$-----------(2)

$\frac{\pi_2}{4}=\pi_3$------------------(3)

and $\pi_1+\pi_2+\pi_3=1$-----------(4)

Solving the above eqns. we get,

Sub (3) in (2),

$$\pi_1+\frac{1}{4}\frac{\pi_2}{4}=\pi_2$$

$$\pi_1=\frac{15\pi_2}{16}$$

(4) becomes,

$$\frac{15\pi_2}{16}+\pi_2+\frac{\pi_2}{4}=1$$

$$\pi_2=\frac{16}{35}$$

$$\pi_1=\frac{15}{16}\times\frac{16}{35}=\frac{3}{7}$$

$$\pi_3=\frac{1}{4}\times\frac{16}{35}=\frac{4}{35}$$

Hence in the long run the probability of buying cereals A, B, C= (π_1,π_2,π_3)

$= (\frac{3}{7},\frac{16}{35},\frac{4}{35})$

3. The transition probability matrix of a Markov chain $\{x_n\}$, n=1,2,3,… having three states 1,2,3 is P= $\begin{pmatrix} 0.1 & 0.5 & 0.4 \\ 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$ and initial distribution is given by $P^{(0)}=(0.7,0.2,0.1)$. Find

(i) P $\{x_2=3/x_0=1\}$

(ii) P $\{x_3=2, x_2=3, x_1=3, x_0=2\}$

(iii) P $\{x_2=3\}$

SOLUTION:

Given tpm is, $P=\begin{pmatrix} 0.1 & 0.5 & 0.4 \\ 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$ and initial distribution $P^{(0)}=(0.7,0.2,0.1)$

(i) $P\{x_2=3/x_0=1\}$

We know that $P_{ij}=P\{x_{n+1}=a_j/x_n=a_i\}$

$P\{x_2=3 / x_0=1\} = P_{13}^{(2)}$

$P^{(1)} = P^{(0)}P$

$=(0.7,0.2,0.1)\begin{pmatrix} 0.1 & 0.5 & 0.4 \\ 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$

$P^{(1)} = (0.22,0.43,0.35)$

$P^{(2)} = P^{(1)} P$

$P^{(2)} = (0.22,0.43,0.35)\begin{pmatrix} 0.1 & 0.5 & 0.4 \\ 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$

$P^{(2)} = (0.385,0.336,0.279)$

$P^{(2)} = 0.279$

(ii) $P\{x_3=2, x_2=3, x_1=3, x_0=2\}$

By the definition of conditional probability

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

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

$P\{x_3=2, x_2=3, x_1=3, x_0=2\}$

$= P\{x_3=2 / x_2=3, x_1=3, x_0=2\}\, P\{x_2=3, x_1=3, x_0=2\}$

$= P\{x_3=2 / x_2=3\}\, P\{x_2=3 / x_1=3, x_0=2\}\, P\{x_1=3, x_0=2\}$

$= P\{x_3=2 / x_2=3\}\, P\{x_2=3 / x_1=3\}\, P\{x_1=3 / x_0=2\}\, P\{x_0=2\}$

$= P_{32} .P_{33} .P_{23} . P_2^{(0)}$

$= 0.4 \times 0.3 \times 0.2 \times 0.2$

$= 0.0048$

(iii) $P\{x_2=3\} = P_3^{(2)} = 0.279.$