UNIT - III

Likelihood Ratio test:-

Neyman-pearson lemma based on the magnitude of the ratio of two probability density functions provides the best test for testing simple hypothesis against simple alternative hypothesis. In this section we shall discuss a general method of test construction called the ~~likelihood~~ Ratio (LR).

properties of Likelihood Ratio test:-

Likelihood ratio (LR) test principle is an intuitive one. If we are testing a simple hypothesis Ho against a simple alternative hypothesis H_1. then the LR principle leads to the same test as given by the neyman pearson lemma.

The LR Test probability of type I error is controlled by suitably choosing the cut off point λ_0. LR Test is generally UMP. If and UMP test at all exists.

(i) under certain conditions $-2\log\lambda$, has an asymptotic chi-square distribution.

(ii) under certain assumptions LR test is consistent.

Test for the mean of a Normal population:-

If the mean of a Normal population has a specified value. Let $(x_1, x_2, \dots x_n)$ be a random sample of size n from the normal population with mean μ and variance σ^2, where μ and σ^2 are unknown.

$$H_0 : \mu = \mu_0 \text{ (specified)} \quad 0 < \sigma^2 < \infty.$$

against the composite alternate hypothesis.

$$H_1 : \mu \neq \mu_0 ; \quad 0 < \sigma^2 < \infty$$

The parameter space Θ is given by.

$$\Theta = [(\mu, \sigma^2) ; \quad -\infty < \mu < \infty, \quad 0 < \sigma^2 < \infty]$$

$$\Theta_0 = [(\mu, \sigma^2) ; \quad \mu = \mu_0, \ 0 < \sigma^2 < \infty]$$

The likelihood function of the sample observations $x_1, x_2, \dots x_n$ is given by

$$L = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right]$$

The maximum likelihood estimates of μ and σ^2 is given by

$$\hat{\mu} = 1/n \sum_{i=1}^{n} x_i = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2$$

$$L(\hat{\Theta}) = \left[\frac{1}{2\pi s^2}\right]^{n/2} \exp(-n/2)$$

The parameter is σ^2 and MLE of σ^2 is $\mu = \mu_0$ is given by

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \mu_0)^2 = S_0^2$$

$$= 1/n \sum (x_i - \bar{x} + \bar{x} - \mu_0)^2$$

$$= 1/n \sum (x_i - \bar{x})^2 + (\bar{x} - \mu_0)^2$$

The product term vanishes, since

$$\sum (x_i - \bar{x})(\bar{x} - \mu_0) = (\bar{x} - \mu_0) \sum (x_i - \bar{x}) = 0$$

$$\hat{\sigma}^2 = s^2 + (\bar{x} - \mu_0)^2 = S_0^2$$

$$L(\hat{\theta}_0) = \left[\frac{1}{2\pi S_0^2}\right]^{n/2} \exp(-n/2)$$

The likelihood ratio criterion.

$$\lambda = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} = \left[\frac{s^2}{S_0^2}\right]^{n/2}$$

$$= \left[\frac{s^2}{s^2 + (\bar{x} - \mu_0)^2}\right]^{n/2}$$

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \quad \text{under } H_0.$$

where $S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{ns^2}{n-1}$

follows student's t distribution with $(n-1)$ d.f thus

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n-1}} \sim t_{n-1}$$

$$\lambda = \frac{1}{\left(1 + \frac{t^2}{n-1}\right)^{n/2}} = \phi(t^2)$$

The critical region of the LR test viz $0 < \lambda < \lambda_0$

$$\left(1+\frac{t^2}{n-1}\right)^{-n/2} \leq \lambda_0$$

$$\left(1+\frac{t^2}{n-1}\right)^{n/2} \geq \lambda_0^{-1} \Rightarrow \frac{t^2}{n-1} \geq (\lambda_0)^{-2/n} - 1$$

$$t^2 \geq (n-1)\left[\lambda_0^{-2/n} - 1\right] = A^2$$

Thus the critical region may be defined

$$|t| = \left|\frac{\sqrt{n}\,(\bar{x}-\mu_0)}{s}\right| \geq A$$

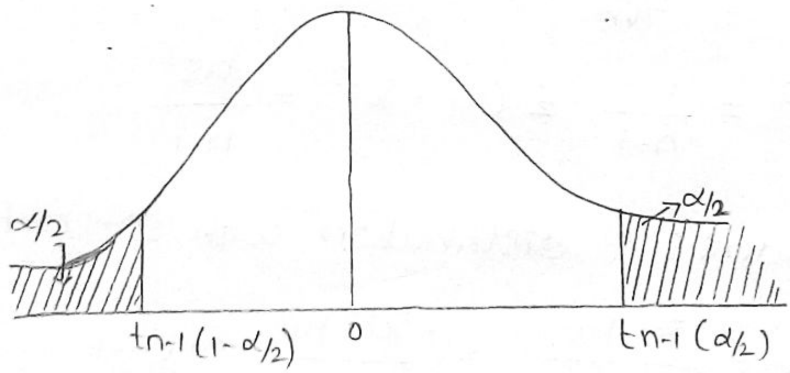
$$P\left[|t| \geq A \mid H_0\right] = \alpha$$

under H_0, the statistic t follows student's t distribution with $(n-1)$ d.F

$$A = t_{n-1}(\alpha/2)$$

$$P\left[t > t_n(\alpha)\right] = \int_{t_n(\alpha)}^{\infty} f(t)\,dt = \alpha$$

where $f(\cdot)$ is the pdf of student's with n pdf



Testing $H_0 : \mu = \mu_0$ against $\mu \neq \mu_0$

$$|t| = \left|\frac{\sqrt{n}\,(\bar{x}-\mu_0)}{s}\right| > t_{n-1}(\alpha/2) \text{ reject } H_0$$

If $|t| < t_{n-1}(\alpha/2)$ H_0 may be accepted.

Test for the equality of means of two normal populations:-

The Independent random variable X_1 and X_2 follows normal distribution $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively.

$$H_0 : \mu_1 = \mu_2 = \mu. \text{ (unspecified)}.$$

$$0 < \sigma_1^2 < \infty, \quad 0 < \sigma_2^2 < \infty.$$

against the alternative hypothesis

$$H_1 : \mu_1 \neq \mu_2 \ \sigma_1^2 > 0, \quad \sigma_2^2 > 0.$$

case-1: population variance are unequal:

$$\Theta = \{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : -\infty < \mu_i < \infty, \ \sigma_i^2 > 0 \quad i=1,2$$

$$\Theta_0 = \{(\mu, \sigma_1^2, \sigma_2^2) : -\infty < \mu < \infty, \ \sigma_i^2 > 0 \quad i=1,2$$

$x_{1i} = (i = 1, 2, \dots m)$ $\quad x_{2j}$ $(j = 1, 2, \dots n)$ be two independent random samples of size m and n from the populations. $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively.

The likelihood function is given by,

$$L = \left(\frac{1}{2\pi\sigma_1^2}\right)^{m/2} \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_{1i} - \mu_1)^2\right] \times \left(\frac{1}{2\pi\sigma_2^2}\right)^{n/2}$$

$$\exp\left[-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n}(x_{2j} - \mu_2)^2\right]$$

The maximum likelihood estimates for $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$

$$\frac{\partial}{\partial \mu_1} \log L = 0 \Rightarrow \hat{\mu}_1 = \frac{1}{M} \sum_{i=1}^{m} x_{1i} = \bar{x}_1$$

$$\frac{\partial}{\partial \mu_2} \log L = 0 \Rightarrow \hat{\mu}_2 = \frac{1}{n} \sum_{j=1}^{n} x_{2j} = \bar{x}_2$$

$$\frac{\partial}{\partial \sigma_1^2} \log L = 0 \Rightarrow \hat{\sigma}_1^2 = \frac{1}{m} \sum_{i=1}^{m} (x_{1i} - \bar{x}_1)^2 = S_1^2$$

and

$$\frac{\partial}{\partial \sigma_2^2} \log L = 0 \Rightarrow \sigma_2^2 = \frac{1}{n} \sum_{j=1}^{n} (x_{2j} - \bar{x}_2)^2 = S_2^2$$

$$L(\hat{\theta}) = \left(\frac{1}{2\pi S_1^2}\right)^{m/2} \left(\frac{1}{2\pi S_2^2}\right)^{n/2} e^{-\frac{(m+n)}{2}}$$

$$L(\theta_0) = \left(\frac{1}{2\pi\sigma_1^2}\right)^{m/2} \exp\left[-\frac{1}{2\sigma_1^2} \sum_{i=1}^{m} (x_{1i} - \mu)^2\right] \times \left(\frac{1}{2\pi\sigma_2^2}\right)^{n/2}$$

$$\exp\left[-\frac{1}{2\sigma_2^2} \sum_{j=1}^{n} (x_{2j} - \mu)^2\right]$$

case - 2: population variances are equal:-

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

$$\theta = \{(\mu_1, \mu_2, \sigma^2) : -\infty < \mu_1 < \infty,\ \sigma^2 > 0\}$$

$$\theta_0 = \{(\mu, \sigma^2) : -\infty < \mu < \infty,\ \sigma^2 > 0\}$$

The likelihood function is given by

$$L = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{(m+n)}{2}} \exp\left[-\frac{1}{2\sigma^2} \left\{\sum_{i=1}^{m} (x_{1i} - \mu_1)^2 + \sum_{j=1}^{n} (x_{2j} - \mu_2)^2\right\}\right]$$

$$L(\hat{\theta}) = \left[\frac{(m+n)}{2\pi\ (ms_1^2 + ns_2^2)}\right]^{\frac{m+n}{2}} \exp\left[-\tfrac{1}{2}(m+n)\right]$$

$$L(\theta_0) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{m+n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{m}(x_{1i}-\mu)^2 + \sum_{j=1}^{n}(x_{2j}-\mu)^2\right\}\right]$$

$$\operatorname{Log} L(\theta_0) = C - \frac{m+n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_i (x_{1i}-\mu)^2 + \sum_j (x_{2j}-\mu)^2\right]$$

where C is constant independent of μ and σ^2.

$$L(\hat{\theta}_0) = \left\{\frac{(m+n)}{2\pi\ (ms_1^2 + ns_2^2 + \frac{mn}{m+n}(\bar{x}_1 - \bar{x}_2)}\right\}^{\frac{m+n}{2}}$$

$$\times \exp\left(\frac{-m+n}{2}\right)$$

$$\lambda = \left\{\frac{1}{1 + \frac{mn(\bar{x}_1 - \bar{x}_2)^2}{(m+n)\ (ms_1^2 + ns_2^2)}}\right\}^{\frac{m+n}{2}}$$

under the null hypothesis $H_0: \mu_1 = \mu_2$ the statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/m + 1/n}}$$

$$s^2 = \frac{1}{m+n-2}\ (ms_1^2 + ns_2^2)$$

$$\lambda = \left[1 + \frac{t^2}{m+n-2}\right]^{-\frac{m+n}{2}}$$

$H_0: \mu_1 = \mu_2\ ;\ \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$

$H_1: \mu_1 \neq \mu_2\ :\ \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$

$$|t| = \left|\frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/m + 1/n}}\right| > t_{m+n-2}(\alpha/2)$$

otherwise H_0 may be accepted.

Test for Equality of variance of two normal populations:-

Consider two normal population $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ where the mean μ_1 and μ_2 and variances σ_1^2 σ_2^2 are unspecified.

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ (unspecified) with μ_1 and μ_2

$H_1 : \sigma_1^2 \neq \sigma_2^2$ μ_1 and μ_2 unspecified.

$$L = \left(\frac{1}{2\pi\sigma_1^2}\right)^{m/2} \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_{1i}-\mu_1)^2\right] \times \left(\frac{1}{2\pi\sigma_2^2}\right)^{n/2}$$
$$\exp\left[-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n}(x_{2j}-\mu_2)^2\right]^2$$

$$L(\hat{\theta}) = \left(\frac{1}{2\pi S_1^2}\right)^{m/2} \cdot \left(\frac{1}{2\pi S_2^2}\right)^{n/2} \exp\left[-\tfrac{1}{2}(m+n)\right]$$

where S_1^2 and S_2^2 are as defined.

The likelihood function.

$$L(\theta_0) = \left[\frac{1}{2\pi\sigma_2}\right]^{\frac{m+n}{2}} \cdot \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_i (x_{1i}-\mu_1)^2 + \sum_j (x_{2j}-\mu_2)^2\right\}\right]$$

and the MLE for μ_1 and μ_2 and σ^2 are given by

$$\hat{\mu}_1 = \bar{x}_1 \quad \text{and} \quad \hat{\mu}_2 = \bar{x}_2$$

$$\hat{\sigma}_2 = \frac{1}{(m+n)}\left[\sum_i (x_{1i}-\hat{\mu}_1)^2 + \sum_j (x_{2j}-\hat{\mu}_2)\right]$$

$$= \frac{1}{m+n}\left[\sum_i (x_{1i}-\bar{x}_1)^2 + \sum_j (x_{2j}-\bar{x}_2)^2\right]$$

$$\hat{\sigma}^2 = \frac{mS_1^2 + nS_2^2}{m+n}$$

$$L(\hat{\theta}_0) = \left[\frac{m+n}{2\pi\,(ms_1^2+ns_2^2)}\right]^{\left(\frac{m+n}{2}\right)} \exp\left[-\tfrac{1}{2}(m+n)\right]$$

$$\lambda = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$$

$$=(m+n)^{\frac{m+n}{2}} \left\{ \frac{(s_1^2)^{m/2}\,(s_2^2)^{n/2}}{\{ms_1^2+ns_2^2\}^{m+n/2}} \right\}$$

$$= \frac{(m+n)^{\frac{m+n}{2}}}{m^{m/2}.\ n^{n/2}} \left\{ \frac{(m s_1^2)^{m/2}\,(n s_2)^{n/2}}{[m s_1^2+ns_2^2]^{\frac{m+n}{2}}} \right\}$$

$$F = \frac{\Sigma(x_{1i}-\bar{x}_1)^2\,/m-1}{\Sigma(x_{2j}-\bar{x}_2)^2\,/n-1} = \frac{S_1^2}{S_2^2}$$

Follows F distribution with $(m-1, n-1)$ d.f.

$$F = \frac{m(n-1)\,s_1^2}{n(m-1)\,S_2^2}$$

$$\left(\frac{m-1}{n-1}\right)F = \frac{ms_1^2}{nS_2^2}$$

$$\lambda = \frac{(m+n)^{\frac{m+n}{2}}}{m^{m/2}\,n^{n/2}} \left\{ \frac{\left(\frac{m-1}{n-1}F\right)^{m/2}}{\left[1+\frac{m-1}{n-1}F\right]^{\frac{(m+n)}{2}}} \right\}$$

$P(F \geq F_2) = \alpha/2$ and

$P(F \geq F_1) = 1-\alpha/2$

Since under H_0 F follows Snedecors F distribution with $(m-1, n-1)$ dF.

$$F_2 = F_{m-1, n-1}(\alpha/2)$$

$$F_1 = F_{m-1, n-1}(1-\alpha/2)$$

$F_{m,n}(\alpha)$ is the upper α point F distribution with (m,n) dF. we have a two tailed F test the critical region being given by

$$F > F_{m-1, n-1}(\alpha/2) \text{ and } F < F_{m-1, n-1}(1-\alpha/2).$$

UNIT - IV

properties of sequential probability ratio test:-

(i) If the sequential probability ratio test of strength (α, β) and the boundary points (A, B) terminates with probability 1 then

$$A \leq \frac{1-\beta}{\alpha} \quad , \quad B \geq \frac{\beta}{1-\alpha} \longrightarrow ①$$

let w_m be the region $B \leq R_i < A$, $i = 1, 2, \ldots m-1$ and $R_m \geq A$ the probability of rejecting H_0 when it is true is

$$\alpha = \sum_{m=1}^{\infty} \int_{w_m} P(x_1 \ldots x_m) / H_0) \, dv^{(m)}$$

$$\leq \sum_{m=1}^{\infty} \int_{w_m} A^{-1} p(x_1 \ldots x_m) H_1) \, dv^{(m)}$$

$$= A^{-1}(1-\beta)$$

where $dv^{(m)} = dx_1 \ldots dx_m$ which establishes 1st inequality in eqn ① similarly the second inequality follows.

ii) If for the choice

$$A = \frac{1-\beta}{\alpha} \quad , \quad B = \frac{\beta}{1-\alpha} \longrightarrow ②$$

$$\alpha' \leq \frac{\alpha}{1-\beta} \quad , \quad \beta' \leq \frac{\beta}{1-\alpha} \text{ and } (\alpha' + \beta') \leq \alpha + \beta \longrightarrow ③$$

$$(\alpha' + \beta') \leq \alpha + \beta \longrightarrow ④$$

$$\frac{1-\beta}{\alpha} \leq \frac{1-\beta'}{\alpha'} \quad , \quad \frac{\beta}{1-\alpha} \geq \frac{\beta'}{(1-\alpha')}$$

$$\alpha' \leq (1-\beta') \frac{\alpha}{1-\beta} < \frac{\alpha}{1-\beta}$$

Similarly $\beta' \le \beta/(1-\alpha)$

$$\frac{1-\alpha}{\beta'} \geq \frac{1-\alpha}{\beta} \Rightarrow \frac{1-\alpha'-\beta'}{\beta} \geq \frac{1-\alpha-\beta}{\beta}$$

$$\frac{1-\beta'}{\alpha'} \geq \frac{1-\beta}{\alpha} \Rightarrow \frac{1-\alpha'-\beta'}{\alpha'} \geq \frac{1-\alpha-\beta}{\alpha}$$

Hence if $(1-\alpha-\beta) > 0$ which is true when α and β are small.

$$\frac{\alpha'+\beta'}{1-\alpha'-\beta'} \le \frac{\alpha+\beta}{1-\alpha-\beta} \Rightarrow \alpha'+\beta' \le \alpha+\beta.$$

So that when α and β are small α' and β' of close to atleast one of the equalities $\alpha' \le \alpha$, $\beta' \le \beta$ is true α' and $\beta' \le \alpha+\beta$. a most straight test then the correct choice of A and B.

ii) suppose the successive observation $(x_1, x_2 \ldots)$ are independent and identically distributed.

$\log [P(x/H_1) / P(x/H_0)]$ where $P[\cdot/H_0]$ and $P[\cdot/H_1]$ are the probability density of a sigle observation x under H_0 and H_1.

a) $P(n < \alpha) = 1$, that is the SPRT eventually terminates.

b) $E(e^{th}) < \alpha$ for $-\alpha < t < t_0$, where $t_0 > 0$ result from which the observations are drawn provided only $P(|z(x)| > 0) > 0$.

let $z_i = \log [P(x_i/H_1) / P(x_i/H_0)]$ then $(z_1, z_2 \ldots)$ is a sequence of i.i.d random variables.

$b = \log B < (z_1 + \cdots + z_i) = S_i < \log A = a \quad i = 1, 2, \ldots$

Hence $T_i = S_{ik} - S_{i-1} < (|b| + |a|) = c \quad i = 1, 2, \ldots, r$

which implies that $p(n > m) \leq p(|T_i| < c, \; i = 1, 2, \ldots r)$

$= p[|T_i| < c]^T$. since T_i are independent

$\lim_{m \to \infty} p(n > m) = \lim_{r \to \infty} [p|T| < c]^r$ keeping k fixed.

$= 0$, if $p(|T| < c) < 1$

Since $p(|z| > 0) > 0$, there exists a constant such that

$p(z > b)$ and $p(z < -h)$ if > 0 non

$p(|T| > c) = p[|z_1 + \ldots z_k| > c]$

Sequential Analysis:-

we have seen that in Neyman-pearson theorem of testing of hypothesis, n the sample size is regarded as a fixed constant and keeping α fixed, we minimize β. But in the sequential analysis theory propounded by A. n the sample size is not fixed but is regarded as a r.v where as both α and β are fixed constant.

Sequential probability Ratio test:-

The best known procedure in sequential testing is the sequential probability Ratio test developed by A.

suppose we want to test the hypothesis $H_0: \theta = \theta_0$

against the alternative $H_1 : \theta = \theta$ for a distribution with pdf $f(x, \theta)$ for any positive Integer m. The likelihood function of a sample $x_1, x_2, \ldots x_m$ from the population with p.d.f (p.m.f) $f(x, \theta)$ is given by.

$$L_{1m} = \prod_{i=1}^{m} f(x_i, \theta_1) \text{ when } H_1 \text{ is true and by}$$

$$L_{0m} = \prod_{i=1}^{m} f(x_i, \theta_0) \text{ when } H_1 \text{ is true and the}$$

likelihood ratio λ_m is given by

$$\lambda_m = \frac{L_{1m}}{L_{0m}} = \frac{\prod_{i=1}^{m} f(x_i, \theta_1)}{\prod_{i=1}^{m} f(x_i, \theta_0)}$$

$$P = \prod_{i=1}^{m} \frac{f(x_i, \theta_i)}{f(x_i, \theta_1)}$$

The SPRT for testing H_0 against H_1 is defined as follows λ_n $(m = 1, 2, \ldots)$ is computed.

(i) if $\lambda_m \geq A$, we terminate the process with the rejection of H_0

ii) If $\lambda_m \leq B$, we terminate the process with acceptance of H_0.

iii) If $B < \lambda_m < A$ we continue sampling by taking an additional observation.

Here A and B $(B < A)$ are the constants which are determined by the relation.

$$A = \frac{1-\beta}{\alpha}, \quad B = \frac{\beta}{1-\alpha} \longrightarrow (3)$$

where α and β are the probability of type I error and type II error respectively.

It is much convenient to deal with $\log \lambda m$ rather than λm.

$$\log \lambda m = \sum_{i=1}^{m} \log \frac{f(x_i, \theta_1)}{f(x_i, \theta_0)} = \sum_{i=1}^{m} z_i \longrightarrow (4)$$

where

$$z_i = \log \frac{f(x_i, \theta_1)}{f(x_i, \theta_0)} \longrightarrow (5).$$

Interms of z_i's SPRT is defined as follows.

(i) if $\Sigma z_i \geq \log A$. reject H_0

ii) If $\Sigma z_i \leq \log B$, roeject H_1 (Accept H_0)

iii) If $\log B < \Sigma z_i < \log A$. by taking an additional observation.

Efficiency of SPRT:-

Let the sequence of observations $(x_1, x_2, \ldots)$ be independent and identically distributed and let the probability densities of a single observations x under H_0 and H_1. be $P[\cdot / H_0]$ and $P(\cdot / H_1)$.

$$z(x) = \log \frac{P(x/H_1)}{P(x/H_0)} \quad a = \log A \text{ and } b = \log B.$$

under the conditions we shall show that SPRT, is superior to a fixed sample size test in the sense that the Average sample number (ASN).

(i) The SPRT terminates with probability both under H_0 and H_1.

ii) It is shown that unless $Z(x) = 1$ with probability $E[Z(x) \mid H_0] < 0$ which implies

$P[Z < 0 / H_0] > 0$ (or) $P(|Z| > 0 / H_0) > 0$.

iii) let $E(|Z| / H_0) < \infty$ and $E[|Z| / H_1] < \infty$.

The approximate expressions for the ASN under H_0 / H_1.

$$E[n / H_0] = \frac{b(1-\alpha) + a\alpha}{E(Z / H_0)} \longrightarrow (1)$$

$$E[n / H_1] = \frac{b\beta + a(1-\beta)}{E(Z / H_1)} \longrightarrow (2)$$

(α, β) is the strength of the SPRT.

$$E[n / H_0] \, E[Z / H_0] = E(S_n / H_0) \longrightarrow (3)$$

$$E(S_n / H_0) = P(S_n \leq b) \, E(S_n / S_n \leq b) + P(S_n \geq a) \, E(S_n / E_n > 0)$$

$$= (1-\alpha) b + \alpha a \longrightarrow (4)$$

For any sequential test procedure which terminates with probability 1.

$$E(n/H_0) \geq \frac{(1-\alpha)\log \beta/1-\alpha + \alpha \log \frac{1-\beta}{\alpha}}{E(Z/H_0)}$$

$$E(n|H_0) = \frac{(1-\alpha)b + \alpha a}{E(Z/H_0)}$$

where a and b have the approximate values.

$$a = \log\frac{1-\beta}{\alpha}, \quad b = \log\frac{\beta}{1-\alpha}$$

$$E(n/H_0)\, E(Z/H_0) = E(S_n/H_0)$$

and compute $E(S_n|H_0)$ for any sequential test

$$E(S_n|H_0) = p\,(H_0 \text{ is rejected})\; E(S_n|H_0)\; H_0 \text{ is (accepted)}$$

$$+\, p\,(H_0 \text{ is rejected})\; E(S_n|H_0)\; H_0 \text{ is rejected.}$$

$$= (1-\alpha)\; E(S_n|H_0)\; H_0 \text{ is accepted} + \alpha E(S_n|H_0 \; H_0 \text{ is rejected})$$

By Jensen's inequality.

$$E(S_n|H_0)\; H_0 \text{ is accepted} \leq \log\; E(e^{S_n}/H_0)\; H_0 \text{ is accepted.}$$

$$= \log\frac{1}{1-\alpha} \sum_{1}^{\infty} \int_{w_m} e^{S_m}\, p(x_1/H_0) \dots p(x_m/H_0)\, dx^m$$

$$e^{S_m} = \frac{p(x_1/H_1) \dots p(x_m/H_0)}{p(x_1/H_0) \dots p(x_m/H_0)}$$

$$\log\frac{1}{1-\alpha} \sum_{1}^{\infty} \int_{w_m} p(x_1/H_1) \dots p(x_m/H_1)\, dv^{(m)}$$

$$= \log\frac{\beta}{1-\alpha}$$

Similarly $E(S_n | H_0)$, H_0 is rejected $< \log[(1-\beta)/\alpha]$

$$E(S_n | H_0) \leq (1-\alpha)\log \frac{\beta}{1-\alpha} + \alpha \log \frac{1-\beta}{\alpha}$$

The fundamental identity of sequence Analysis:-

Lemma - 1::

Let z be a random variables such that

(a) $p(z>0) > 0$ and $p(z<0) > 0$

(b) $\phi(t) = E(e^{tz})$ exists for any real value t and

(c) $E(z) \neq 0$.

Then there exists $E(z) < 0$ then $c > 0$ and if $E(z) > 0$ then $c < 0$.

Since $p(z>0) > 0$ there exists such that.

$p(z>c) = \delta > 0$ Hence.

$$\phi(t) = E(e^{tz}) = \int e^{tz}\, dF$$

$$\geq \int_{z>c} e^{tz}\, dF > e^{tc}\, p(z>c) \text{ if } t>0 \rightarrow (1)$$

furthermore $\phi(0) = 1$ and the slope of $\phi(t)$

$\phi'(F) = E(ze^{tx})$ at $t=0$ is $E(z)$ If $E(x) \leq 0$.

similarly when $E(z) > 0$, there exists $a z < 0$ such that

$\phi(z) = 1$.

Since $\phi^n(t) = E(z^x e^{tz})$, the condition of implies that $\phi^n(t) > 0$ for all t, which inturn implies that it can be one minimum at the almost. Hence here is only one value z of t other than 0 at which $\phi(t) = 1$ for if $\phi(t) = 1$ at $\bar{c}_1, \bar{c}_2$ besides zero. the atleast two minima are implied.

Lemma - 2 :-

let $x_1, x_2, \ldots$ be a sequence of i.i.d observations on a random variable x with probability density $p(\cdot | H_1)$ under a hypothesis H. consider any sequential procedure with a given stopping true and let n be the no. of. observations needed for coming to a decision.

it is $z_i = z(x_i$ is any measurable function then

$$p(n < \alpha | H) = 1 \Rightarrow$$

$$E\{e^{tS_n}[\phi(t)]^{-n} | H\} = P(n < \infty, H_1)$$

where under H_t the probability density of x is

$$P[\cdot | H_t] = \frac{e^{tz(\cdot)}}{\phi(t)} \; P(\cdot | H)$$

$S_n = Z_1 + \ldots + Z_n$ and $\phi(t) = E(e^{tz} | H)$

let W_m denote the region in R^m leading to the termination of the sequential procedure at the m^{th} stage then.

$$E\{e^{tsn}[\phi(t)]^{-n} \mid H\} = \sum_{1}^{\infty} \int_{w_m} [\phi(t)]^{-mt} e^{sm}$$

$$P(X/H) \cdots P(X_m \mid H)\, dv^{(m)}$$

$$\Rightarrow \sum_{i}^{\infty} \int_{w_m} P(X_1 \mid H_t) \cdots P(X_m \mid H_t)\, dv^{(m)}$$

$$= P(n < \infty \mid H_t)$$

The fundamental identity of sequential analysis

consider the SPRT based on a sequence of i.i.d observation and let z be as defined the logarithm of the ratio. under the hypothesis H_0 and H_1: If H_1 is any hypothesis such that.

$$P(|z| > 0 \mid H) > 0. \quad \text{then}$$

$$E\{e^{tsn}(\phi(t))^{-n} / H\} = 1$$

which is called the fundamental identify of sequential analysis.

It is easy to see that $P[|z| > 0 \mid H] > 0$

If $P(|z| > 0 / H_t) > 0$. where H_t is defined in eqn.

Hence by $P[n < \infty / H_t] = 1$. The result.

UNIT - $\underline{V}$

Non parametric Methods:-

Most of the statistical tests that we discussed so far had the following two features in common.

i) The form of the frequency function of the parent population from which the samples have been drawn is assumed to be known, and

ii) They were concerned with testing statistical hypothesis about the parameters of this frequency function or estimating its parameters.

For example, almost all the exact sample tests of significance are based on the fundamental assumption that the parent population is normal and are concerned with testing or estimating the means and variance of these population. Such tests, which deal with the parameters of the population are known as parametric tests. Thus the parametric statistical test is a test whose model specifies certain conditions about the parameter of the population from which the samples are drawn.

On the other hand, a Non-parametric (NP) test is a test does not depend on the particular form of this basic frequency function from which the samples are drawn. In other words, non-parametric test does not make any assumption regarding the form of the population.

However certain assumptions associated with N.P. Tests are.

Sample observations are independent

The variable under study is continuous

P.d.f is continuous

Lower order moments exist.

Obviously these assumptions are fewer and much weaker than those associated with parametric tests.

Advantages and Disadvantages of N.P. Methods over parametric methods :-

Below we shall give briefly the comparative study of parametric and non-parametric methods and their relative merites and demerits.

<u>Advantages of N.P. Methods :-</u>

i) N.P. Methods are readily comprehensible, very simple and easy to apply and donot require complicated sample theory.

ii) No assumptions is made about the form of the frequency function of the parent population from which sampling is done.

iii) No parametric technique will apply to the data which are mere classification, while Np methods exist to deal with such data.

iv) Since the socio-economic data are not in general normally distributed. Np tests have found applications in psycometry, sociology and Educational statistics.

v) N.P. Tests are available to deal with the data which are given in ranks or whose seemingly numerical scores have the strength of ranks. For instance, non parametric test can be applied of the scores are given in grades such as A, B ..etc.

Disadvantages of N.P. Tests:

(i) Np tests can be used only in the measurements are nominal methods. Even in that case, if a parametric test can exist it is more powerful than the parametric test. In other words, if all the assumptions of a statistical model are ~~eakk~~ satisfied the data and if the measurements are of required strength then the N.p. test wasteful of time and data

ii) so far no N.p. methods exist for testing interactions in Analysing variance model unless special assumptions about the additivity of the model model.

iii) N.p. test are designed to test statistical hypothesis only and estimating the parameters.

Basic distribution :-

Let z be a continuous random variable with a p.d.f $f(\cdot)$. Let $z_1, z_2, \ldots z_n$ be a random sample of size n from $f(\cdot)$ and let $x_1, x_2, \ldots x_n$ be the corresponding ordered sample. Then the Joint density of $x_1, x_2, \ldots x_n$ is given by

$$g(x_1, x_2, \ldots x_n) = n!\ f(x_1)\, f(x_2) \ldots f(x_n), \quad -\infty < x_1 < x_2 < \ldots < x_n < \infty$$

$$U_i = \int_{-\infty}^{x_i} f(z)\, dz = F(x_i) \quad (i=1, 2, \ldots n)$$

when $F(\cdot)$ is the distribution function of z. Since $F(x_i)$ is a uniform random variable on $[0, 1]$ U_i $(i=1, 2, \ldots n)$ are random variables following uniform distribution on $[0, 1]$.

Thus the Joint density $K(\cdot)$ of the random variables u_i $(i=1,2,\ldots,n)$

$$K(u_1, u_2 \ldots u_n) = n! \quad 0 \leq u_1 \leq u_2 < \ldots < u_k \leq 1$$

and does not depend on $f(\cdot)$.

$$E(U_i) = \int_0^1 \cdots \int_0^{u_1} \int_0^{u_2} u_i n_1 \, du_1 du_2 \ldots du_n$$

$$= \frac{i}{n+1}$$

$$E(u_i) - E(u_{i+1}) = \frac{i}{n+1} - \frac{i-1}{n+1} = \frac{1}{n+1}$$

which is independent of $f(\cdot)$.

Wald - Wolfowitz Run Test:-

Definition :-

A run is defined as a sequence of letters of one kind surrounded by a sequence of letters of the other kind, and the number of elements in a run is usually referred to as the length(l) of the run.

we have in order a run of x $(l=2)$, a run of y $(l=3)$ a run of x $(l=1)$, a run of y $(l=1)$ etc..

Test for Randomness:-

Another application of the 'run' theory is in testing the randomness of a given set of observations. Let $x_1, x_2, \ldots x_n$ be the set of observations arranged in the order in which they occur, i.e x_i is the ith observations in the outcome of experiment. Then for each of the observations, and write A if the observation is above and B if it is below the median value. Thus we get a sequence of A's and B's of the type.

A B B A A A B A B B

under the null hypothesis H_0 that the set of observation is random, the number of run U in r.v. with

$$E(U) = \frac{n+2}{2} \text{ and}$$

$$var(U) = \frac{n}{4}\left(\frac{n-2}{n-1}\right)$$

Median Test:-

Median test is a statistical procedure for testing if two independent ordered sample differ in their central tendencies. In other words, it gives information if two independent samples are likely to have been drawn from the population with same median.

As in 'run' test, let $x_1, x_2, \dots x_{n_1}$ and $y_1, y_2, \dots y_{n_2}$ be two independent ordered samples from the populations with pdf's $f_1(\cdot)$ and $f_2(\cdot)$ respectively. The measurements must be at least ordinal. Let $z_1, z_2, \dots z_{n_1} \cdot n_2$ be the combined ordered sample. Let m_1 be the number of x's and m_2 the number of y.

under $H_0: f_1(\cdot) = f_2(\cdot)$, the Joint distribution of m_1 and m_2 is the hypergeometric distribution with probability function.

$$p(m_1, m_2) = \frac{\binom{n_1}{m_1}\binom{n_2}{m_2}}{\binom{n_1+n_2}{m_1+m_2}}$$

$$\sum_{m_1=1}^{m_1} p(m_1, m_2) = \alpha$$

The distribution of m_1 under H_0 is also Hyper geometric

$$E(m_1) = \frac{n_1}{2}, \text{ if } N = n_1 + n_2 \text{ is even}$$

$$= \frac{n_1}{2} \cdot \frac{N-1}{N} \text{ if } N \text{ is odd}$$

$$\text{var}(m_1) = \frac{n_1 n_2}{4(N-1)} \text{ if } N \text{ is even}$$

$$= \frac{n_1 n_2 (N+1)}{4N^2} \text{ if } N \text{ is odd.}$$

$$Z = \frac{m_1 - E(m_1)}{\sqrt{\text{var}(m_1)}} \sim N(0,1) \text{ asymptotically.}$$

Sign Test :-

consider a situation where it is desired to compare two things or materials under various sets of conditions. An experiment is thus conducted under the following circumstances.

(i) where there are pairs of observations on two things being compared.

ii) For any given period, each of the two observations is made under similar extraneous conditions.

iii) Different pairs are observed under different conditions.

This implies that the differences $d_i = x_i - y_i$; $i = 1, 2, \dots n$ have different variances and thus renders the paired t-test invalid which would have otherwise been unless was obvious non-normality.

The only assumptions are;

(i) Measurements are such that the deviations $d_i = x_i - y_i$, can be expressed in terms of positive or negative signs.

ii) variables have continuous distribution

iii) d_i's are independent.

procedure :-

Let (x_i, y_i) $i = 1, 2, \ldots n$ be n paired sample observations drawn from the two populations with pdf's $f_1(\cdot)$ and $f_2(\cdot)$.

Null Hypothesis $H_0 : f_1(\cdot) = f_2(\cdot)$

To test H_0 consider $d_i = x_i - y_i$ $(i = 1, 2, \ldots n)$

$$P[X - Y > 0] = 1/2 \text{ and } P(X - Y < 0) = 1/2$$

$$u_i = \begin{cases} 1, & \text{if } x_i - y_i > 0 \\ 0, & \text{if } x_i - y_i < 0 \end{cases}$$

u_i is Bernoulli variate with $p = P(x_i - y_i) > 0 = 1/2$

since u_i's $i = 1, 2, \ldots n$ are independent

$U = \sum_{i=1}^{n} u_i$, the total number of positive deviations, is a Binomial variate with parameters n and p $(= 1/2)$. the Number of positive deviations be K.

$$P(U \leq K) = \sum_{r=0}^{K} \binom{n}{r} p^r q^{n-r}, \quad (p = q = 1/2 \text{ under } H_0)$$

$$= (1/2)^n \sum_{r=0}^{K} \binom{n}{r} = p'$$

If $p' \leq 0.05$, we reject H_0 at 5% level of significance and if $p' > 0.05$ we conclude that the data do not provide any evidence against the null hypothesis.

For large Samples $n \geq 30$

$$E(U) = np = n/2$$

$$var(U) = npq = n/4$$

$$Z = \frac{U - E(U)}{\sqrt{var(U)}}$$

$$Z = \frac{U - n/2}{\sqrt{(n/4)}}$$ is asymptotically $N(0,1)$