INFERENTIAL STATISTICS

SYLLABUS

UNIT - I

UNIT-I

POPULATION:

The word population in statistics is used to refer to any collection of information about an individual which can be numerically specified.

For example: Weights of persons, heights of persons, price of a particular commodity etc.,

The population may be finite or infinite.

Finite:

A population contains a finite number of values is called a finite population.

For example: Number of students in a classroom, No. of. pages in a book.

Infinite:

A population which contains an infinite number of values is called a infinite population.

For example: Number of stars in the sky, No. of. Leaves in a tree, etc.,

SAMPLE:

A Small part selected from a population is called a sample.

For example: A handful of rice will access (find) the quality of pack of rice.

Sample size:

The number of individual in a sample is called the sample size.

For example: No. of. Courses in a college

Parameter:

Any statistical constant obtained from a population is called the parameter.

For example: $\mu \rightarrow$ The population mean; $\sigma^2 \rightarrow$ The population variance.

Statistic:

Any statistical constant obtained from a sample is called a statistic.

For example: $\bar{x}$ - sample mean; $S^2 \rightarrow$ sample variance.

Parameter space:

The set of all possible values of a an unknown parameters is called parameter space

H → [Captheta]

For example: If X ~ N(μ, σ^2), then the parameter space is defined as,

H = {(μ, σ^2): -∞ < μ < ∞ ; 0< σ < ∞}

Statistical Inference:

Statistical inference refers to the process of using a sample statistic to draw a valid inference or conclusion about a population parameter. There are two types of problems in statistical inference. Namely, (i) Estimation, (ii) Test of hypothesis.

Estimation:

Introduction: Given a random samplex_1, x_2, x_3, x_4.....x_n of size 'n' from a population with p.d.f f(x, θ) but with known parameterθ. Our problem is to find and estimate for θ in terms of the sample values. In order to estimate the value of unknown parameter we may gives an procedure termed as estimation.

Definition: Estimation is defined as a process by which sample information is used to estimate the numerical values of once or more parameter of the population.

Estimator and Estimate:

A function of sample value is called an estimator. While its numerical value is called an estimate . thus the values of parameter is to be estimated is called estimator and the value of the sample is used to estimate the value of the parameter is called the estimate.

For example: $\bar{x}$ is an estimator of population mean μ and other hand, if the $\bar{x}$ =25, for a sample, the estimate of population mean μ is equal to 25.

(ie) μ = 25 → μ= $\bar{x}$, the sample mean.

The estimation may be classified into two types. They are (i) Point estimation

(ii) Interval estimation.

Point estimation:

Point estimation is a process in which a single statistics like mean, median, and standard deviation etc, is used as estimation of an population parameter. Thus, point estimation is defined as a single number which represents the estimate of a population parameter.

Interval Estimation:

Interval estimation is also a process in which, it is possible to estimate an interval within which the values of parameter is expected to lie. The estimated interval is termed as confidence Interval.

Best Estimate:

The best estimate would be one that falls nearest to the true value of the parameter to be estimated.

Characteristics of Good Estimator or Best Estimator:

An estimator satisfies the following conditions:

(i) Unbiasedness (ii) Consistency (iii) Efficiency (Iv) Sufficiency

Unbiasedness:

An estimator T_n = T ($x_1, x_2, x_3, x_4 \ldots\ldots\ldots\ldots x_n$) is said to be an unbiased estimator of a population parameter $\vartheta(\theta)$ if

E(T_n) , for all $\theta \in$ H In general E(statistics) = parameter.

Remarks:

(i) If E(T_n) $> \theta$, T_n is said to be positively biased.
(ii) If E(T_n) $< \theta$, T_n is said to be negatively biased.
(iii) The amount of biased b(θ) $being\ given\ by$ b(θ) = E(T_n) - $\vartheta(\theta)$, $\theta \in$ H
(iv) If E(T_n) $\neq \theta$, it is said to be biased estimator of θ.

PROBLEM :1

$x_1, x_2, x_3, x_4 \ldots\ldots\ldots\ldots x_n$ is a random sample from a normal population N(μ, 1).

Show that t= $\frac{1}{n}\sum_{i=1}^{n} x_i^2$, is an unbiased estimator of 1+μ^2.

SOLUTION:

We have to prove that, E(t) = 1+ μ^2 , We are given that $X_i \sim$ N(μ, 1)

E(x_i) = μ ; V(x_i)=1 $\forall$ i=1,2, 3,.....n

Now, V(x_i) = E(x_i2) – [$E\ (x_i)$]^2

E(x_i)^2 = V(x_i) + [$E\ (x_i)$]^2

E(x_i)^2 = 1+ μ^2 , N(μ, 1)

$\therefore$ Consider t = $\frac{\sum_{i=1}^{n} x_i^2}{n}$

E(t) = E$\left[\frac{\sum_{i=1}^{n} x_i^2}{n}\right]$

= $\frac{1}{n}\sum_{i=1}^{n} E(x_i^2)$

= $\frac{1}{n}\sum_{i=1}^{n}(1 + \mu^2)$

= $\frac{1}{n}$. n$(1 + \mu^2)$

E(t) = $1+ \mu^2$ Hence t is an unbiased estimator of $1+ \mu^2$

PROBLEM: 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) = θ

Also V(T) = E(T^2) – [E(T)]^2

V(T) = E(T^2) - θ^2 $\therefore$V(T) $>$ 0

E(T^2) – V(T) + θ^2

Since E(T^2) $\neq$ θ^2

T^2 is a biased estimator for θ^2.

PROBLEM: 3

Show that $\frac{[\sum x_i\ (\sum x_i - 1)]}{n(n-1)}$ is an unbiased estimate of θ^2, for the sample x_1, x_2, x_3, x_4.....x_n drawn on X which takes the values 1 or 0 with respective probabilities θ And (1- θ).

SOLUTION: Since x_1, x_2, x_3, x_4.....x_n is a random sample from Bernoulli population with parameter θ , T = $\sum_{i=1}^{n} x_i \sim B(n, \theta)$

E(T) = nθ; V(T) = nθ (1- θ).

$\therefore$ E $\left\{\frac{\sum x_i\ (\sum x_i - 1)}{n(n-1)}\right\}$ = $\left\{\frac{T(T-1)}{n(n-1)}\right\}$

$= \frac{1}{n(n-1)}$ E [T(T-1)]

$= \frac{1}{n(n-1)}$ {E(T^2) – [E(T)} $\because$ V(T) = E(T^2) – [E(T)]

$= \frac{1}{n(n-1)}$ {V(T) = [E(T)]^2 – [E(T)} E(T^2) = V(T) + [E(T)]

$= \frac{1}{n(n-1)}$ { nθ (1- θ) + $n^2\theta^2$ − nθ}

$= \frac{1}{n(n-1)}$ { nθ - nθ^2 + $n^2\theta^2$ − nθ}

$= \frac{1}{n(n-1)}$ [n$\theta^2(n-1)$] = θ^2.

Hence $\frac{\sum x_i\ (\sum x_i\ -1)}{n(n-1)}$ is an unbiased estimator of θ^2.

PROBLEM: 4

Prove that a random sampling from the normal population. The sample mean is an unbiased estimator of the population mean μ.

SOLUTION:

Let $x_1, x_2, x_3, x_4 \ldots\ldots\ldots\ldots x_n$ be a random sample of size n from N(μ, σ^2).

(Ie) $X_i \sim$ N(μ, σ^2) E(x_i)= μ and V(x_i) = σ^2

$\bar{x} = \frac{\sum x_i}{n}$ E($\bar{x}$) =E[$\frac{\sum x_i}{n}$] = $\frac{1}{n}\sum_{i=1}^{n} E(x_i)$

$= \frac{1}{n}$ [E(x_1) + $E(x_2)$ + ……..+$E(x_n)$].

$= \frac{1}{n}[\mu + \mu + \cdots + \mu]$

$= \frac{1}{n}$ .n μ = μ

$\therefore$ The sample mean $\bar{x}$ is an unbiased estimator of population mean μ.

PROBLEM: 5

For the normal population, prove that the sample variance is not an unbiased estimator of population variance σ^2.

SOLUTION:

Let $x_1, x_2, x_3, x_4 \ldots\ldots\ldots\ldots x_n$ be a random sample of size ‘n’ from N(μ, σ^2).

(Ie) $X_i \sim N(\mu, \sigma^2)$ $\forall$ i=1,2,3,.....n

$E(x_{i,}) = \mu$ and $V(x_i) = \sigma^2$

Consider, the sample variance $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})2$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu + \mu - \bar{x})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{(x_i - \mu)^2 + (\mu - \bar{x})^2 + 2(x_i - \mu)(\mu - \bar{x})\}$$

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

Consider , $\sum_{i=1}^{n}(x_i - \mu) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu$ $\bar{x} = \frac{\sum x_i}{n}$

$= n\bar{x} - n\mu$ $n\bar{x} = \sum x_i$

$= n(\bar{x} - \mu)$

$= -n(\bar{x} - \mu)$

Now put $\sum_{i=1}^{n}(x_i - \mu) = -n(\mu - \bar{x})$ in equation 1 we get

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 + (\mu - \bar{x})^2 + \frac{2}{n}[-n(\mu - \bar{x}).(\mu - \bar{x})]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 + (\mu - \bar{x})^2 - 2((\mu - \bar{x})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 + (\mu - \bar{x})^2$$

Taking expectation on both sides,

$E(S^2) = E\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 + (\mu - \bar{x})^2\}$

$= E\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\} - \{E(\mu - \bar{x})^2\}$

$= \frac{1}{n}\sum_{i=1}^{n} E(x_i - \mu)^2 - E(\mu - \bar{x})^2$

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

$= \frac{1}{n} n\sigma^2 - V(\bar{x})$

$= \sigma^2 - \frac{\sigma^2}{n}$ $= \sigma^2$ (1-1/n)

$= \sigma^2$ (n-1/n)

$= (\frac{n-1}{n}).\ \sigma^2 \neq \sigma^2$

Hence the sample variance is not an unbiased estimator of the population variance σ^2.

CONSISTENCY:

An estimator T_n = T(x_1, x_2, x_3, x_4.....x_n), based on a random sample of size n, is said to be consistent estimator of $\vartheta(\theta), \theta \in$ H , the parameter space, if T_n converges to $\vartheta(\theta)$ in probability , (i.e) if $T_n \xrightarrow{p} \vartheta(\theta)$ as n→ ∞. In other word's, T_n is a consistent estimator of $\vartheta(\theta)$

If for every $\varepsilon > 0, \eta > 0$, there exist a positive integer n $\geq m$ (ε, η) such that

P{ | $T_n - \vartheta(\theta)| < \varepsilon$} → 1

P{ | $T_n - \vartheta(\theta)| < \varepsilon$} → 1- η ; ∀ $n \geq m$. where m is some very large value of n.

Consistency as well as unbiased estimator

PROBLEM:

Prove that in sampling from a N (μ, σ^2) population, the sample mean is a consistent estimator of μ.

Solution:

In sampling from a N(μ, σ^2)population, the sample mean $\bar{x}$ is also normally distributed as N($\mu, \sigma^2/n$).

(i.e) E($\bar{x}$) = μ and V($\bar{x}$) = σ^2/n

Let $x_i \sim$ N (μ, σ^2), I = 1, 2,…..n Thus , we have

E(x_i)= μ and V (x_i) = σ^2

Consider the sample mean $\bar{x}$,

$\bar{x} = \frac{\sum x_i}{n}$

Taking Expectation on both sides, E ($\bar{x}$) =E[$\frac{\sum x_i}{n}$]

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

$= \frac{1}{n}$ [E(x_1) + $E(x_2)$ + ……..+$E(x_n)$].

$= \frac{1}{n} [\mu + \mu + \cdots + \mu]$

$$= \frac{1}{n} \cdot n\mu = \mu$$

$$V(\bar{x}) = V\left(\frac{\sum x_i}{n}\right)$$

$$= \frac{1}{n^2} V(\sum x_i)$$

$$= \frac{1}{n^2} \{V(x_1) + V(x_2) + \cdots \ldots \ldots + V(x_n)\}$$

$$= \frac{1}{n^2} [\sigma^2 + \sigma^2 + \sigma^2 + \cdots \ldots + \sigma^2$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$V(\bar{x}) = \frac{\sigma^2}{n}$$

Thus, $\lim_{n\to\infty}$,

$$\lim_{n\to\infty} E(\bar{x}) = \mu$$

$$\lim_{n\to\infty} V(\bar{x}) = \frac{\sigma^2}{n} = 0$$

Thus E $(\bar{x}) = \mu$, $V(\bar{x}) = 0$

Hence, the sample mean $\bar{x}$ is a consistent estimator of population mean(μ).

Consistency as well as unbiased estimator

PROBLEM:

Prove that the sample median is normal population is the consistent estimator of the population mean 'μ'.

Solution: Let $X_i \sim$ N (μ, σ^2), i= 1,2,……n

Then, $\bar{X} \sim$ N $(\mu, \sigma^2/n)$ and the median $\tilde{X} \sim$ N $(\mu, \frac{\pi}{2}, \sigma^2/n)$

$$E(\tilde{X}) = \mu \text{ and } V(\tilde{X}) = \frac{\pi}{2}, \sigma^2/n$$

As, $\lim_{n\to\infty} E(\widetilde{X}) = \lim_{n\to\infty} \mu = \mu$

$$\lim_{n\to\infty} V(\widetilde{X}) = \lim_{n\to\infty} \frac{\pi}{2}, \sigma^2/n = 0$$

Thus, as $n \to \infty$ $E(\widetilde{X}) = \mu$ & $V(\tilde{X}) = 0$

Hence, the median in normal population is the consistent estimator of the population mean μ.

INVARIANCE PROPERTY OF CONSISTENT ESTIMATOR:

If T_n is a consistent estimator of $\gamma(\theta)$ and $\psi[\gamma(\theta]$ is a continuous function of $\gamma(\theta)$, then $\psi(T_n)$ is a consistent estimator of $\psi[\gamma(\theta)]$.

Proof:

Since T_n is a consistent estimator of $\gamma(\theta)$,

$T_n \to \gamma(\theta)$ as $n \to \infty$ i.e., for every $\xi > 0, \eta > 0$, a positive integer n≥ m ((ε, η)

Such that P

$P\{|T_n - \gamma(\theta)| < \varepsilon\} > 1- \eta$, $\forall$ $n \geq m$ → (1) Since $\psi(.)$ is a continuous function, for every $\xi > 0$, however small, ∃ a positive number ε , such that $|\Psi(T_n) - \Psi(\gamma(\theta)| < \varepsilon$, whenever $|T_n - \gamma(\theta)| < \varepsilon$

(i.e) $|T_n - \gamma(\theta)| < \varepsilon \Rightarrow |\Psi(T_n) - \Psi(\gamma(\theta)| < \varepsilon$, → (2)

For two events A and B , if A ⇒B, then

A $\underline{C}$ B ⇒ P(A) ≤ P(B) or P(B) ≥ P(A) → (3)

From (2) and (3) , we get $P\{|\Psi(T_n) - \Psi\gamma(\theta)| < \varepsilon_1\} \geq$P $[|T_n - \gamma(\theta)| < \varepsilon]$

⇒ $P\{|\Psi(T_n) - \Psi\gamma(\theta)| < \varepsilon_1\} \geq 1- \eta$; $\forall$ $n \geq m$ [using (1)]

⇒ $\Psi(T_n) \xrightarrow{p} \Psi(\gamma(\theta)]$, as n→ ∞ or $\Psi(T_n)$ is a consistent estimator of $\gamma(\theta)$.

Properties of Consistent estimator:

1. The Consistency would give increasing with increasing the size of the sample.

2. A Consistent estimator is unbiased I the limit but an unbiased estimator may or may not be consistent estimator.

3. The property of Consistency is a limiting property.

4. If there exists one consistent statistic we can construct an infinite number of consistent statistics.

Since each of the statistic converges in probability to the parameter.

Theorem:

Let $\{T_n\}$ be the sequence of estimator of 'θ'. Such that

i) E $(T_n) \to \theta'$ as n$\to \infty$

ii) E $(T_n) \to \theta$ as n$\to \infty$

Then T_n is a consistent estimator of θ'.

Solution:

Let $\varepsilon > 0$ by Tchebecheve's inequality

We have, P[$|T_n$ – E$(T_n) < \varepsilon$] $\geq 1 - \frac{V(T_n)}{\varepsilon^2}$

Taking limit as n$\to \infty$

$$\lim_{n\to\infty} P\left[|T_n - E(T_n)| < \varepsilon\right] \geq \lim_{n\to\infty} 1 - \frac{V(T_n)}{\varepsilon^2}\Big|$$

i.e) $\lim_{n\to\infty} P\left[|T_n - \theta')| < \varepsilon\right] \geq 1 - \frac{0}{\varepsilon^2}\Big|$

$\lim_{n\to\infty}$ P[$|T_n - \theta')| < \varepsilon$] ≥ 1 |

$\Rightarrow T_n$ is a consistent estimator of θ'

(i.e) $T_n \xrightarrow{p} \theta'$ as n$\to \infty$.

Efficiency:

If T_1 is the most efficient estimator with variance V_1 and T_2 is any other estimator with varianceV_2, then the efficiency E of T_2 is defined as:

$$E = \frac{V_1}{V_2}$$

Obviously, E cannot exceed unity. E = $\frac{V_1}{V_2} < 1$.

More efficient:

Let V(T_1) and V(T_2) be the variance of the estimator T_1 and T_2 respectively. Then the estimator T_1 is said to be more efficient then the estimator T_2 .

If V(T_1) < V(T_2).

Example: If X $\sim$ N (μ, σ^2) Then the mean $\bar{X} \sim$ N $(\mu, \sigma^2/n)$ and the median

$\tilde{X} \sim$ N $(\mu, \frac{\pi}{2}.\sigma^2/n)$ Here, V($\bar{X}$) = σ^2/n , V($\tilde{X}$) = $\frac{\pi}{2}.\sigma^2/n$

Hence V($\bar{X}$) is less than V($\tilde{X}$). Thus, the samples mean $\bar{X}$ is more efficient, and then the sample median is normal results.

Problem: 1

Consider the normal population N (μ, σ^2) the sample mean is the efficient of the population meanμ.

Solution:

In this case Var (sample mean) = σ^2/n

Var (sample median) = $\frac{\pi}{2}.\sigma^2/n$

E = $\frac{V(sample\ mean)}{V(sample\ median)}$

= $\frac{\sigma^2/n}{\frac{\pi}{2}.\sigma^2/n}$

= $\frac{\sigma^2}{n}$ x $\frac{2n}{\pi\sigma^2}$

= $\frac{2}{\pi}$ = $\frac{2}{22/7}$ $\Rightarrow$ 2 x $\frac{7}{22}$ = $\frac{7}{11}$ = 0.64

E = 0.64

$\therefore$ E < 1

Thus, the sample mean is more efficient then sample median.

Similarly, the sample mean is more efficient than any other estimator. Hence the sample mean is the most efficient estimator of the population mean μ for the normal population N (μ, σ^2).

Problem: 2

A random sample (X_1, X_2, X_3, X_4, X_5) of size '5' is drawn from a normal population with unknown mean 'μ'. Consider the following estimators to estimate μ.

i) $t_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$ ii) $t_1 = \frac{X_1 + X_2}{2} + X_3$

iii) $t_3 = \frac{2X_1 + X_2 + \lambda X_3}{3}$, where λ is such that t_3 is unbiased estimator of 'μ'

iv) Find λ, t_1 and t_2 unbiased.

v) State giving reasons, the estimator which is best among t_1 ,t_2 and t_3.

Solution:

Given that , E(x_i) = μ, i = 1,2,3,4,5

i) Consider, E (t_1) = $E\left\{\frac{X_1+X_2+X_3+X_4+X_5}{5}\right\}$

$= \frac{1}{5}\{E(X_1)+E(X_2)+E(X_3)+E(X_4)+E(X_5)\}$

$= \frac{1}{5}\{\mu+\mu+\mu+\mu+\mu\}$

$= \frac{5\mu}{5}$

E (t_1) = μ

t_1 is an unbiased estimator of the parameter μ.

ii) E (t_2) = $E\left\{\frac{X_1+X_2}{2}+X_3\right\}$

$= \frac{1}{2}E(X_1+X_2)+E(X_3) = \frac{1}{2}\{E(X_1)+E(X_2)+E(X_3)\}$

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

$= \mu+\mu$

$= 2\mu \neq \mu$

Hence it is biased estimator of μ.

iii) Given that, t_3 is unbiased estimator of μ.

(i.e) E(t_3) = μ $\rightarrow$ (1)

E (t_3) = $E\left\{\frac{2X_1+X_2+\lambda X_3}{3}\right\}$

$= \frac{1}{3}\{E(2X_1)+E(X_2)+E(\lambda X_3)\}$

$= \frac{1}{3}\{2E(X_1)+E(X_2)+\lambda E(X_3)\}$

$= \frac{1}{3}\{2\mu+\mu+\lambda\mu\}$

E (t_3) = $\frac{1}{3}\{3\mu+\lambda\mu\} \rightarrow$ (2)

From (1) and (2), we can write

$\mu = \frac{1}{3}\{3\mu+\lambda\mu\}$

$3\mu = 3\mu+\lambda\mu$

$3\mu - 3\mu = \lambda\mu$

$\lambda\mu = 0$

$\lambda = 0$

$t_3 = \frac{2X_1+X_2+0}{3}$ = $t_3 = \frac{2X_1+X_2}{3}$

$E(t_3)$ =E[$\frac{2X_1+X_2}{3}$] = $\frac{1}{3}$ {2E (X_1) + $E(X_2)$}

= $\frac{1}{3}$ [2μ + μ] = $\frac{3\mu}{3}$

$E(t_3)$ = μ

t_3 is an unbiased estimator of μ.

i) Consider, V(t_1) = V$\left\{\frac{X_1+X_2+X_3+X_4+X_5}{5}\right\}$

= $\frac{1}{25}$ {V (X_1)+$V(X_2)$ + $V(X_3)$ + $V(X_4)$+$V(X_5)$}

= $\frac{1}{25}$ { $\sigma^2 + \sigma^2 + \sigma^2 + \sigma^2 + \sigma^2$}

= $\frac{5\sigma^2}{25} = \frac{\sigma^2}{5}$ = (0.2) σ^2

V(t_1) = (0.2) σ^2

V (t_2) = V$\left\{\frac{X_1+X_2}{2} + X_3\right\}$

= $\frac{1}{4}$ {V (X_1) + $V(X_2)$} + $V(X_3)$

= $\frac{1}{4}$ { $\sigma^2 + \sigma^2$} + σ^2

= $\frac{2\,\sigma^2}{4}$ + σ^2

= $\frac{\sigma^2}{2}$ + σ^2 = $\frac{3\,\sigma^2}{2}$ = (1.5) σ^2

V (t_3) = V$\left\{\frac{2X_1+X_2}{3}\right\}$

= $\frac{1}{9}$ {V $(2X_1 + X_2)$

= $\frac{1}{9}$ {4 V(X_1) + V(X_2)}

= $\frac{1}{9}$ [4$\sigma^2 + \sigma^2$]

= $\frac{5\,\sigma^2}{9}$

= (0.56) σ^2

By comparing variance of t_1 ,t_2 and t_3 having the minimum variance.

Thus t_1 is best estimate of the parameter than t_2 and t_3.

Cramer Rao Inequality :(meaning)

Cramer rao inequality provides a lower bound on the variance of an unbiased of the parameter. If an estimator t_n is unbiased. Then, it is also efficient estimator if and only if. The variance of estimator t_n attains the Cramer rao lower bound. In this case it must be a sufficient estimator of the parameters.

Thus, the ratio of the Cramer rao lower bound to the actual variance of any unbiased estimator for a parameter is called the efficiency of that estimator.

Cramer Rao Inequality: (Theorem)

Statement: Let 'X' be a continuous r.v with p.d.f f(x, θ) and likelihood function 'L'. Let't' be an unbiased estimator of some function of θ say $\gamma(\theta)$.

Then, V (t) $\geq \frac{\left[\frac{d}{d\theta}\gamma(\theta)\right]^2}{E\left\{\frac{\partial}{\partial\theta}\log L\right\}^2}$ (or) V (t) $\geq \frac{[\gamma'(\theta)]^2}{I(\theta)}$, Type equation here.

Where I (θ) = $E\left\{\frac{\partial}{\partial\theta}\log L\right\}^2$

Proof: Given that t' is an unbiased estimator of$\gamma(\theta)$.

Let t = t ($x_1, x_2, x_3, \ldots\ldots\ldots x_n$) be an unbiased estimator of $\gamma(\theta)$.

Such that, E(t) = $\gamma(\theta)$

$\gamma(\theta)$ = E(t) = $\int t.L\,dx$ [$\because$ L is joint pdf of x_i 's i= 1,2,3,....n]

$\gamma(\theta) = \int t.L\,dx$

P differentiate with respect to , we get

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

Multiply and divide by L $\because$ log L = $\frac{1}{L}$

$\gamma'(\theta) = \int t\ \frac{1}{L}\left(\frac{\partial L}{\partial\theta}\right) L\,dx$ L= L(x, θ) = $\prod_{i=1}^{n} f(x_i, \theta)$

$= \int t\ \left(\frac{\partial L}{\partial\theta}\right) L\,dx$ $\int L(x,\theta)dx = 1$

$\gamma'(\theta)$ = E $\left(t.\frac{\partial \log L}{\partial\theta}\right)$ $\rightarrow$ (1)

$\therefore$ L be the likelihood function of the r.v 'X' and also be a joint density function. Then, we have $\int L dx = 1$ p.D.w.r.to θ and using Since L is the joint pdf of

$(x_1, x_2, x_3, \ldots\ldots\ldots x_n)$ regularity conditions given above,

We get $\int \frac{\partial L}{\partial \theta}$ dx = 0

Multiply and divided by L

$$\int t \left(\frac{1}{L}\frac{\partial L}{\partial \theta}\right) L \, dx = 0$$

$$\int t \left(\frac{\partial \log L}{\partial \theta}\right) L \, dx = 0$$

$$E\left(\frac{\partial \log L}{\partial \theta}\right) = 0 \quad \rightarrow (2)$$

Now Consider, $E\left(t, \frac{\partial \log L}{\partial \theta}\right) = \gamma'(\theta)$ Cov(X,Y) = E(XY) – E(X).E(Y)

$$Cov\left(t.\frac{\partial \log L}{\partial \theta}\right) = E\left(t, \frac{\partial \log L}{\partial \theta}\right) - E(t) \cdot E\left[\frac{\partial \log L}{\partial \theta}\right]$$

$$Cov\left(t.\frac{\partial \log L}{\partial \theta}\right) = E\left(t, \frac{\partial \log L}{\partial \theta}\right) - 0 \quad [\text{using (2)}]$$

$$Cov\left(t.\frac{\partial \log L}{\partial \theta}\right) = \gamma'(\theta) \quad (\text{from (1)})$$

$$Cov\left(t.\frac{\partial \log L}{\partial \theta}\right) = \gamma'(\theta) \quad \rightarrow (3)$$

For two variables say x & y we have , r

$$r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \leq 1 \quad (\because r \leq 1)$$

from this $Cov(x,y) \leq \sigma_x \sigma_y$

$[Cov(xy)]^2 \leq \sigma_x{}^2 \sigma_y{}^2 \quad \Rightarrow [Cov(xy)]^2 \leq V(x).V(y)$

Now, $\{Cov\left(t.\frac{\partial \log L}{\partial \theta}\right)\}^2 \leq V(t) \cdot V\left(\frac{\partial \log L}{\partial \theta}\right)$

$$[\gamma'(\theta)]^2 \leq V(t) \cdot V\left(\frac{\partial \log L}{\partial \theta}\right) \rightarrow (4) \quad (\text{from(3)})$$

Consider,

$$V\left(\frac{\partial \log L}{\partial \theta}\right) = E\left\{\left(\frac{\partial \log L}{\partial \theta}\right)^2\right\} - \left[E\left(\frac{\partial \log L}{\partial \theta}\right)\right]^2$$

$$= E\left(\frac{\partial \log L}{\partial \theta}\right)^2 - 0 \qquad \because E\left(\frac{\partial \log L}{\partial \theta}\right) = 0$$

$$V\left(\frac{\partial \log L}{\partial \theta}\right) = E\left(\frac{\partial \log L}{\partial \theta}\right)^2 \rightarrow (5)$$

Using (5), (4) become

$$[\gamma'(\theta)]^2 \leq V(t) \cdot E\left(\frac{\partial \log L}{\partial \theta}\right)^2$$

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

$$V(t) \geq \frac{[\gamma'(\theta)]^2}{E\left(\frac{\partial \log L}{\partial \theta}\right)^2}$$ hence proved.

Problem:

Let $x_1, x_2, x_3, \ldots\ldots\ldots x_n$ denotes a random sample from a Poisson distribution has the mean $\theta > 0$. It is known that y = $\sum x_i$, i= 1,2,3,.....n is sufficient for θ.

Show that $\frac{y}{n} = \bar{x}$ is an efficient statistic for θ.

Solution:

Given that, $x_i \sim p(\theta)$, then p(x) = $\frac{e^{-\theta}\theta^x}{x_!}$

Taking log on both sides,

Log p(x) = $-\theta + x\log\theta - \log x_i$

P. D. w. r to θ

$$\frac{\partial \log p(x)}{\partial \theta} = -1 + \frac{x}{\theta} \Rightarrow \frac{-\theta + x}{\theta} \Rightarrow \frac{x-\theta}{\theta}$$

$$E\left\{\frac{\partial \log L}{\partial \theta}\right\}^2 = E\left\{\left(\frac{x-\theta}{\theta}\right)^2\right\}$$

$$= \frac{1}{\theta^2} E(x - \theta)^2$$

$$= \frac{1}{\theta^2} E[x - E(X)]^2 \qquad V(X) = E[X - E(X)]^2$$

$$= \frac{1}{\theta^2} V(x) \qquad \text{Poisson distribution}$$

$$= \frac{\theta}{\theta^2} \Rightarrow \frac{1}{\theta} \qquad \text{Mean } E(x) = \theta$$

$$E\left\{\frac{\partial \log L}{\partial \theta}\right\}^2 = \frac{1}{\theta} \qquad \text{Variance } V(x) = \theta$$

∴ Cramer Rao – Lower bound is given by

$\frac{1}{n\,E\left(\frac{\partial\, logL}{\partial\theta}\right)^2} = \frac{1}{n\left(\frac{1}{\theta}\right)} = \theta/n = V(\bar{x})$ $\quad\because\ V(\bar{x}) = \frac{\theta}{n}$

$\bar{x}$ is an efficient statistic for θ.

Sufficiency:

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 ($x_1, x_2, x_3, \ldots\ldots x_n$) is an estimator of a parameter θ, based on a sample $x_1, x_2, x_3, \ldots\ldots x_n$ of size n from the population with density f(x, θ) such that the conditional distribution of $x_1, x_2, x_3, \ldots\ldots x_n$ given T, is independent of θ, then T is sufficient estimator for θ.

Factorization Theorem (Neymann):

The necessary and sufficient condition for a distribution to admit sufficient statistics is provided by the 'Factorization theorem' to Neymann.

Statement: T = t(x) is sufficient for θ if and only if the joint density function L (say), of the sample values can be expressed in the form:

L = g_θ [t(x)]. h(x)

Where (as indicated) g_θ[t(x)] depends on θ and x only through the value of t(x) and h(x) is independent of θ.

Problem: 1

Let $x_1, x_2, x_3, \ldots\ldots x_n$ be a random sample drawn from a population with pdf f(x, θ) = $\theta\, x^{\theta-1}$; $0 < x < 1, \theta > 0$.

(i) Show that $t_n = \prod_{i=1}^{n} x_i$ is sufficient estimator for θ.

Solution: Consider, the likelihood function,

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

$$= \theta^n \frac{x_1^{\theta}}{x_1}, \frac{x_2^{\theta}}{x_2}, \ldots\ldots\ldots\ldots \frac{x_n^{\theta}}{x_n}$$

$$= \theta^n \prod_{i=1}^{n} x_i^{\theta} \cdot \frac{1}{\prod_{i=1}^{n} x_i}$$

L = g(t, θ) = h($x_1, x_2, x_3 \ldots\ldots x_n$)

Where h($x_1, x_2, x_3 \ldots\ldots x_n$) = 1, g($t_n, \theta$) = $\theta^n \prod_{i=1}^{n} x_i^{\theta} \cdot \frac{1}{\prod_{i=1}^{n} x_i}$

$\therefore t_n = \prod_{i=1}^{n} x_i$ is sufficient estimator for 'θ'.

Problem: 2

Obtain the sufficient statistic of parameter λ of Poisson distribution.

Solution:

Let $x_1, x_2, x_3 \ldots\ldots\ldots x_n$ be a random sample of size 'n' from Poisson distribution with parameter λ . Thus, we have

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

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

$$= \prod_{i=1}^{n} P(x_i) = P(x_i), p(x_2), p(x_3), \ldots\ldots p(x_n)$$

$$= \frac{e^{-\lambda}\lambda^x}{x!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdot \frac{e^{-\lambda}\lambda^{x_3}}{x_3!} \ldots\ldots\ldots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$

$$= \frac{(e^{-\lambda})^n \lambda^{x_1+x_2+x_3+\cdots\ldots+x_n}}{\prod_{i=1}^{n} x_i!}$$

$$= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} = e^{-n\lambda}\lambda^{\sum x_i} \cdot \frac{1}{\prod_{i=1}^{n} x_i!}$$

$L = g(t, \lambda) \cdot h(x)$

Where $g(t, \lambda) = e^{-n\lambda}\lambda^{\sum x_i}$

$h(x) = \prod_{i=1}^{n} x_i !$

$t = \sum x_i$ is a sufficient estimator of the parameter .

Problem: 3

Let $x_1, x_2, x_3 \ldots\ldots\ldots x_n$ be a random sample from a uniform distribution with pdf $f(x, \theta) = \frac{1}{\theta}$ $0 < x < \theta, \theta > 0$. Obtain the sufficient statistic of the parameterθ.

Solution:

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

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

$$= \left(\frac{1}{\theta}\right) \cdot \left(\frac{1}{\theta}\right) \cdot \left(\frac{1}{\theta}\right) \ldots\ldots\ldots\ldots \left(\frac{1}{\theta}\right)$$

$= \left(\frac{1}{\theta}\right)^n$

$L \neq g(t, \theta) \cdot h(x)$

Now consider the order statistic, Let $x_{(1)}, x_{(2)}, x_{(3)} \ldots\ldots\ldots x_{(n)}$ be an 'n' independent order samples from the uniform distribution.

Thus, $\leq x_{1,} \leq x_{2,} \leq \cdots \ldots \ldots \ldots \leq x_n \leq \theta.$

$$x_{(n)} \leq \theta$$

Hence the largest sample $x_{(n)}$ is related with the parameterθ.

$\therefore x_n$ Is sufficient statistic or sufficient estimator of the parameter θ.

Properties of Sufficient estimator or sufficient statistic:

1. A sufficient estimator is always consistent.
2. A sufficient estimator is most efficient. If an efficient estimator exist.
3. A sufficient estimator may or may not be unbiased.

Rao Blackwell Theorem:

Let X and Y be random variables such that

$E(y) = \mu$ and $V(y) = \sigma_y{}^2 > 0$

Let E(Y/X=x) = ∅(x). Then (i) E [∅(X)] = μ and (ii) Var[∅(X)] ≤ Var(Y).

Proof:

Let f(x, y) be the joint pdf of random variables X and Y, $f_1(.)$ and $f_2(.)$

The marginal p.d.f's of X and Y respectively and h(y/x) be the conditional p.d.f of y for given X=x such that h(y/x) = { f(x, y)/ $f_1(x)$}

Now, we consider

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

$= \int_{-\infty}^{\infty} y.\frac{f(x,y)}{f_1(x)} dy$ s

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

We have E[Y/X=x] = ∅(x)

$\therefore \frac{1}{f_1(x)} \int_{-\infty}^{\infty} y.f(x,y)dy = \emptyset(x) \rightarrow (2)$

$$\int_{-\infty}^{\infty} y.f(x,y)dy = \emptyset(x).f_1(x)$$

From (1), we observe that the conditional distribution of Y given X=x does not depends on the parameter μ.

Hence X is sufficient statistic for μ. Also

$\therefore$ E{ $\emptyset(x)$} = E {E(y/x)} = E(y) = μ

E[$\emptyset(x)$] = μ $\rightarrow$ (3)

This establishes part (i) of the theorem.

Now V(Y) = E(Y^2) – [E(Y)]^2

= E{ [Y - E(Y)]^2} = E {(Y- μ)^2} $\because$ E(y) = μ

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

= E {(Y- $\emptyset(x)$)^2 + [$\emptyset(x) - \mu$)]^2 + 2 [(Y- $\emptyset(x)$. ($\emptyset(x)$ – μ]}

V(Y) = E {(Y- $\emptyset(x)$)^2 +E [$\emptyset(x) - \mu$)]^2 + 2E [(Y- $\emptyset(x)$. ($\emptyset(x)$ – μ]}

= E {(Y- $\emptyset(x)$)^2 }+E [$\emptyset(x) - \mu$)]^2} + 0 $\because$ E(y- $\emptyset(x)$ =0

= E {(Y- $\emptyset(x)$)^2 } + E { [$\emptyset(x) - E(\emptyset(x))$]^2} $\because$ E($\emptyset(x)$ = μ

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

V [$\emptyset(x)$] $\leq$ V (Y) hence proved.

UNIT-II

UNIT – II

METHODS OF ESTIMATION:

So far we have been discussing the requesting of a good estimator. Now we shall briefly outline some of the important methods for obtaining such estimators. Commonly used methods are.

i) Method of maximum likelihood estimation (MLE)
ii) Method of moments
iii) Method of minimum variance
iv) Method of least squares
v) Method of minimum chi-squares
vi) Method of inverse probability

Method of maximum likelihood estimation:

Likelihood function: let $x_1, x_2,...x_n$ be a random sample of size n from a population with density function $f(x,\theta)$. Then the likelihood function of the sample values $x_1, x_2,...x_n$ usually denoted by $L = L(\theta)$ is their joint density function,

Given by, $L = f(x_1,\theta)\, f(x_2,\theta)...f(x_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$

L gives the relative likelihood that the r.v's assume a particulars set of values

$x_1, x_2,...x_n$. For a given sample $x_1, x_2,...x_n$, L becomes a function of the variable θ, the parameter.

The principle of maximum likelihood consists in finding an estimator for the unknown parameter $\theta = (\theta_1, \theta_2,... \theta_k)$, say, which maximises the likelihood function $L(\theta)$ for variations in parameter, i.e., we wish to find $\theta = (\theta_1, \theta_2, ...\theta_k)$ so that

$$L(\theta) > L(\theta) \forall \theta \in \Theta, i.e, L(\theta) = \sup L(\theta) \forall \theta \in \Theta$$

Thus if there exists a function $\theta = \theta(x_1, x_2, ...x_n)$ of the sample values which maximises L for variations in θ, then θ is to be taken as an estimator of θ. θ is usually called maximum likelihood estimator (M.L.E). Thus θ is the solution, if any, of $\frac{\partial L}{\partial \theta} = 0 \, and \, \frac{\partial^2 L}{\partial \theta^2} < 0$.

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 θ. The first of the two equations can be rewritten as:

$\frac{1}{L}.\frac{\partial L}{\partial \theta}=0 \Rightarrow \frac{\partial^2 L}{\partial \theta^2}<0$ a from which is much more convenient from practical point of view.

PROPERTIES OF MLE:

1. MLE is not unbiased
2. MLE is not unique.
3. MLE's are consistent.
4. MLE's are always consistent estimators but need not be unbiased.
5. If MLE exists, it is the most efficient in the class such estimators.
6. If a sufficient estimator exists, it is a function of the maximum likelihood estimator.
7. MLE's have the invariance property of MLE. If T is the MLE of θ and $\varphi(\theta)$ is one to one function of θ, then $\varphi(T)$ is the MLE of $\varphi(\theta)$.
8. MLE's are asymptotic normality.

PROBLEM:

In random sampling from Normal population N(μ, σ^2), find the MLE 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:

Given

$X \square N(\mu, \sigma^2), then$

$$f(x_i, \theta)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}; -\infty<x<\infty, -\infty<\mu<\infty$$

$$\sigma^2<0$$

Then the likelihood function of the random sample is given by

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

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

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

Taking log *on both sides*

$$\log L = n\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \to 1$$

i) When σ^2 is known, the likelihood equation for estimating μ is

$$\frac{\partial \log L}{\partial \mu} = 0 \Rightarrow -0 - 0 - \frac{1}{2\sigma^2} 2\sum_{i=1}^{n}(x_i-\mu)(-1)$$

$$\Rightarrow +\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i - n\mu = 0$$

$$\Rightarrow -n\mu = -\sum_{i=1}^{n} x_i \Rightarrow n\mu = \sum_{i=1}^{n} x_i$$

$$\Rightarrow \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$\Rightarrow \hat{\mu} = \bar{x}$$

Hence M.L.E for μ is the sample mean $\bar{x}$.

ii) When μ is known, the likelihood equation for estimating σ^2 is

$$\frac{\partial \log L}{\partial \sigma^2} = 0 \Rightarrow 0 - \frac{n}{2}.\frac{1}{\sigma^2} - \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2\left(\sigma^2\right)^{-1-1}(-1)$$

$$\Rightarrow -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$$

$$\Rightarrow \frac{1}{2\sigma^2}\left[-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2\right] = 0$$

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

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

$$\Rightarrow \sum_{i=1}^{n}\left(x_i - \mu\right)^2 = n\sigma^2$$

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

$$\Rightarrow \sigma^2 = s^2$$

iii) The likelihood equations for simultaneous estimation of μ and σ^2 are:

$$\frac{\partial \log L}{\partial \mu} = 0 \; and \; \frac{\partial \log L}{\partial \sigma^2} = 0$$

$$Thus\ giving\ \hat{\mu} = \bar{x}\ and\ \sigma^2 = \frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{n}$$

$\sigma^2 = s^2$, Sample variance.

Hence, the MLE of (μ, σ^2) is $(\bar{x}, s^2)$

PROBLEM:

Prove that the maximum likelihood estimate of the parameter α of a population having density function: $\frac{2}{\alpha^2}(\alpha - x), 0 < x < \alpha,$ for a sample of unit size is 2x, x being the sample value. Show also that the estimate is biased.

Solution:

For a random sample of unit size (n = 1), the likelihood functions is:

$$L(\alpha) = f(x,\alpha) = \frac{2}{\alpha^2}(\alpha - x); 0 < x < \alpha$$

$$\log L = \log 2 - \log \alpha^2 + \log(\alpha - x)$$

$$\log L = \log 2 - \log \alpha + \log(\alpha - x)$$

$$\frac{\partial \log L}{\partial \alpha} = 0 - \frac{2}{\alpha} + \frac{1}{\alpha - x}$$

$$\frac{\partial \log L}{\partial \alpha} = 0 \Rightarrow - - \frac{2}{\alpha} + \frac{1}{\alpha - x} = 0$$

$$-\frac{2}{\alpha} = -\frac{1}{\alpha - x}$$

$$2(\alpha - x) = \alpha$$

$$2\alpha - 2x = \alpha$$

$$2\alpha - 2x - \alpha = 0$$

$$\alpha - 2x = 0$$

$$\therefore \alpha = 2x$$

$$\text{Consider, } E(\hat{\alpha}) = E(2x) = \int_0^{\alpha} 2x\, f(x)dx = \int_0^{\alpha} 2x . \frac{2}{\alpha^2}(\alpha - x)dx$$

$$= \frac{4}{\alpha^2}\int_0^{\alpha} \left(x\alpha - x^2\right)dx = \frac{4}{\alpha^2}\left[\frac{\alpha x^2}{2} - \frac{x^3}{3}\right]_0^{\alpha} = \frac{4}{\alpha^2}\left[\frac{\alpha^3}{2} - \frac{\alpha^3}{3}\right]$$

$$= \frac{4}{\alpha^2}\left[\frac{3\alpha^3 - 2\alpha^3}{6}\right] = \frac{4}{\alpha^2}\left[\frac{\alpha^3}{6}\right] = \frac{2\alpha}{3}$$

$$E(\hat{\alpha}) = \frac{2\alpha}{3} \neq \alpha$$

Hence $\hat{\alpha} = 2x$ is not an unbiased estimator for α.

PROBLEM:

i) Find the MLE for the parameter λ of a Poisson distribution on the basis of a sample of size n. Also find its variance.

ii) Show that the sample mean $\bar{x}$, is sufficient for estimating the parameter λ of the Poisson distribution.

Solution: The probability function of the poisson distribution with parameter 'λ'.

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

Consider, the likelihood function

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

$$= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} .... \frac{e^{-n\lambda}\lambda^{x_n}}{x_n!}$$

$$\therefore L = \frac{\left(e^{-\lambda}\right)^n \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} = \frac{e^{-n\lambda}\lambda^{\Sigma x_i}}{\prod_{i=1}^{n} x_i!}$$

$$\log L = -n\lambda + \Sigma x_i \log \lambda - \sum_{i=1}^{n} \log(x_i!)$$

$$\therefore \frac{\partial \log L}{\partial \lambda} = -n + \frac{\Sigma x_i}{\lambda} - 0$$

$$\frac{\partial \log L}{\partial \lambda} = 0 \Rightarrow -n + \frac{\Sigma x_i}{\lambda} = 0$$

$$\Rightarrow -n = -\frac{\Sigma x_i}{\lambda}$$

i.e.,

$$\Rightarrow n = \frac{\Sigma x_i}{\lambda}$$

$$\Rightarrow \lambda = \frac{\Sigma x_i}{n}$$

$$\Rightarrow \lambda = \bar{x}$$

Thus the M.L.E for λ is the sample mean $\bar{x}$. The variance of estimate is given by:

$$\frac{1}{V(\hat{\lambda})} = E\left\{-\frac{\partial^2}{\partial \lambda^2}(\log L)\right\} = E\left\{-\frac{\partial}{\partial \lambda}\left(-n + \frac{n\bar{x}}{\lambda}\right)\right\}$$

$$= E\left\{-\left(-\frac{n\bar{x}}{\lambda^2}\right)\right\} = \frac{n}{\lambda^2} E(\bar{x}) = \frac{n}{\lambda}$$

$$\therefore \left[E(\bar{x})\right] = \lambda$$

$$V(\hat{\lambda}) = \frac{\lambda}{n}$$

For the poisson distribution with parameter λ, we have

$$\frac{\partial \log L}{\partial \lambda} = -n + \frac{n\bar{x}}{\lambda} = n\left(\frac{\bar{x}}{\lambda} - 1\right) = P(\bar{x}, \lambda),$$ a function of $\bar{x}$ and λ only.

Hence $\bar{x}$ is sufficient for estimating λ.

PROBLEM:

Let $x_1, x_2, ...x_n$ be a random sample from the uniform distribution with p.d.f

$$f(x,\theta) = \begin{cases} 1/\theta ; 0 < x < \infty, \theta > 0 \\ 0 \quad ; otherwise \end{cases}$$

i) Obtain the MLE for θ.

Solution:

Consider,

$$L = \prod_{i=1}^{n} f(x_i, \theta)$$
$$= f(x_1,\theta), f(x_2,\theta), ... f(x_n,\theta)$$
$$L = \frac{1}{\theta}, \frac{1}{\theta}, ... \frac{1}{\theta} = \left(\frac{1}{\theta}\right)^n = \theta^{-n}$$

Now,

$$\log L = -n \log \theta$$
$$\frac{\partial \log L}{\partial \theta} = \frac{-n}{\theta}$$

Consider, $\frac{\partial \log L}{\partial \theta} = 0 \Rightarrow \frac{-n}{\theta} = 0$

i) $\hat{\theta} = \infty$ it is impossible, so that we consider the order statistics.

ii) Let $x_{(1)}, x_{(2)}, ..., x_{(n)}$ be an order. Random samples of size 'n' from the given population. So that, $0 \le x_{(1)} \le x_{(2)} \le ... \le x_{(n)} \le \theta$. $x_n \le \theta$.

The maximum value of θ is consistent with $x_n \le \theta$. Thus the largest sample observation $x_{(n)}$ is the MLE of θ.

METHODS OF MOMENTS:

Let $f(x;\theta_1,\theta_2,...,\theta_k)$ be the density function of the parent population with k parameters $\theta_1, \theta_2, ..., \theta_k$. If μ_r' denotes the r[th] moment about origin, then

$$\mu_r' = \int_{-\infty}^{\infty} x^r f(x;\theta_1,\theta_2,...,\theta_k) dx, (r = 1,2,...k) \rightarrow *$$

In general $\mu_1', \mu_2', ..., \mu_k'$ will be function of the parameters $\theta_1, \theta_2, ..., \theta_k$.

Let x_i, i = 1,2,...n be a random sample of size n from the given population. The method of moments consists in solving the k – equations(*) for $\theta_1, \theta_2, ..., \theta_k$ in terms of $\mu_1', \mu_2', ..., \mu_k'$ and then replacing these moments $\mu_r'; r = 1, 2, ..., k$ by the sample moments,

$e.g., \theta_i = \theta_i(\hat{\mu}_1', \hat{\mu}_2', ..., \hat{\mu}_k') = \theta_i(m_1', m_2', ..., m_k')$; i = 1,2,...k where m_i' 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 OF METHOD OF MOMENTS:

i) The estimates obtained by the method of moments will have asymptotically normal distribution for large n.
ii) The estimator obtained by this method are less efficient that those obtained from the principle of MLE.
iii) The estimator obtained by this method is consistent.
iv) The estimator obtained by this method is identical estimator if MLE's are obtained as linear functions of the moments.

PROBLEM:1

For the double Poisson distribution:

$$p(x) = p(X = x) = \frac{1}{2}.\frac{e^{-m_1}.m_1^x}{x!} + \frac{1}{2}.\frac{e^{-m_2}.m_2^x}{x!}; x = 0,1,2,...$$

Show that the estimates for m_1 and m_2 by the method of moments are: $\mu_1' \pm \sqrt{\mu_2' - \mu_1' - \mu_1'^2}$

Solution:

$$We\ have, \mu_1' = \sum_{x=0}^{\infty} x - p(x) = \frac{1}{2}\sum_{x=0}^{\infty} x.\frac{e^{-m_1}.m_1^x}{x!} + \frac{1}{2}.\frac{e^{-m_2}.m_2^x}{x!}$$
$$= \frac{1}{2}m_1 + \frac{1}{2}m_2$$

(Since the first and second summations are the means of Poisson distributions with parameters m_1 and m_2 respectively.)

$$\mu_2' = \sum_{x=0}^{\infty} x^2 - p(x) = \frac{1}{2}\left\{\sum_{x=0}^{\infty} x^2 . \frac{e^{-m_1} . m_1^x}{x!} + \sum_{x=0}^{\infty} x^2 . \frac{e^{-m_2} . m_2^x}{x!}\right\}$$

$$= \frac{1}{2}\left\{\left(m_1^2 + m_1\right) + \left(m_2^2 + m_2\right)\right\}$$

$$= \frac{1}{2}\left\{\left(m_1 + m_2\right) + \left(m_1^2 + m_2^2\right)\right\}$$

$$= \frac{1}{2}\left\{2\mu_1' + \left(m_1^2 + m_2^2\right)\right\}$$

$$= \frac{1}{2}\left\{2\mu_1' + m_1^2 + \left(2\mu_1' - m_1\right)^2\right\}$$

$$= \frac{1}{2}\left\{2\mu_1' + m_1^2 + 4\left(\mu_1'\right)^2 + m_1^2 - 4\mu_1' m_1\right\}$$

$$\mu_2' = \mu_1' + m_1^2 + 2\left(\mu_1'\right)^2 - 2\mu_1' m_1$$

$$m_1^2 - 2m_1\mu_1' + \left(2\left(\mu_1'\right)^2 + \mu_1' - \mu_2'\right) = 0$$

$$\therefore \hat{m}_1 = \frac{2\mu_1' \pm \sqrt{4\mu_1'^2 - 4\left(2\mu_1'^2 + \mu_1' - \mu_2'\right)}}{2}$$

$$\therefore \hat{m}_1 = \mu_1' \pm \sqrt{\mu_2' - \mu_1' - \left(\mu_1'\right)^2}$$

Similarly on substituting for m_1 in terms of m_2 we get

$$m_2^2 - 2m_2\mu_1' + \left(2\left(\mu_1'\right)^2 + \mu_1' - \mu_2'\right) = 0$$

Solving for m_2, we get

$$\hat{m}_2 = \mu_1' \pm \sqrt{\mu_2' - \mu_1' - \left(\mu_1'\right)^2}$$

PROBLEM:2

A random variable X takes the values,0,1,2, with respective probabilities $\frac{\theta}{4N} + \frac{1}{2}\left(1 - \frac{\theta}{N}\right), \frac{\theta}{2N} + \frac{\alpha}{2}\left(1 - \frac{\theta}{N}\right)$ and $\frac{\theta}{4N} + \frac{1-\alpha}{2}\left(1 - \frac{\theta}{N}\right)$, where N is a known number and α, θ are unknown parameters. If 75 independent observations on x yielded the values 0,1,2 with frequencies 27, 38, 10 respectively, estimate θ and α by the method of moments.

Solution:

$$\mu_1^{'} = E(x) = \Sigma x.p(x)$$

$$= 0.\left[\frac{\theta}{4N} + \frac{1}{2}\left(1-\frac{\theta}{N}\right)\right] + 1.\left[\frac{\theta}{2N} + \frac{\alpha}{2}\left(1-\frac{\theta}{N}\right)\right] + 2\left[\frac{\theta}{4N} + \frac{1-\alpha}{2}\left(1-\frac{\theta}{N}\right)\right]$$

$$= 0 + \frac{\theta}{2N} + \frac{\alpha}{2}\left(1-\frac{\theta}{N}\right) + \frac{2\theta}{4N} + \frac{2(1-\alpha)}{2}\left(1-\frac{\theta}{N}\right)$$

$$= \frac{\theta}{2N} + \frac{\alpha}{2}\left(1-\frac{\theta}{N}\right) + \frac{\theta}{2N} + (1-\alpha)\left(1-\frac{\theta}{N}\right)$$

$$= \frac{2\theta}{2N} + \left(1-\frac{\theta}{N}\right)\left[\frac{\alpha}{2} + (1-\alpha)\right]$$

$$= \frac{\theta}{N} + \left(1-\frac{\theta}{N}\right)\left[\frac{\alpha+2-2\alpha}{2}\right] \Rightarrow \frac{\theta}{N} + \left(1-\frac{\theta}{N}\right)\left[\frac{2-\alpha}{2}\right]$$

$$= \frac{\theta}{N} + \left(1-\frac{\theta}{N}\right)\left(1-\frac{\alpha}{2}\right) \Rightarrow \frac{\theta}{N} + \left[1-\frac{\theta}{N}-\frac{\alpha}{2}+\frac{\theta}{N}.\left(\frac{\alpha}{2}\right)\right]$$

$$= 1-\frac{\alpha}{2} + \frac{\theta}{N}.\left(\frac{\alpha}{2}\right)$$

$$\mu_1^{'} = 1-\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right) \to 1$$

$$\mu_2^{'} = E(x^2) = \Sigma x^2 p(x)$$

$$= \theta^2.\left[\frac{\theta}{4N} + \frac{1}{2}\left(1-\frac{\theta}{N}\right)\right] + 1^2\left[\frac{\theta}{2N} + \frac{\alpha}{2}\left(1-\frac{\theta}{N}\right)\right] + 2^2.\left[\frac{\theta}{4N} + \frac{1-\alpha}{2}\left(1-\frac{\theta}{N}\right)\right]$$

$$= \frac{\theta}{2N} + \frac{\alpha}{2}\left(1-\frac{\theta}{N}\right) + \frac{4\theta}{4N} + \frac{4(1-\alpha)}{2}\left(1-\frac{\theta}{N}\right)$$

$$= \frac{\theta}{2N} + \frac{\theta}{N} + \left(1-\frac{\theta}{N}\right)\left[\frac{\alpha}{2} + 2(1-\alpha)\right]$$

$$= \frac{3\theta}{2N} + \left(1-\frac{\theta}{N}\right)\left[\frac{\alpha+4-4\alpha}{2}\right] \Rightarrow \frac{3\theta}{2N} + \left(1-\frac{\theta}{N}\right)\left[\frac{4-3\alpha}{2}\right]$$

$$= \frac{3\theta}{2N} + \left(1-\frac{\theta}{N}\right)\left(2-\frac{3\alpha}{2}\right) \Rightarrow \frac{3\theta}{2N} + 2 - \frac{2\theta}{N} - \frac{3\alpha}{2} + \frac{3\alpha}{2}\left(\frac{\theta}{N}\right)$$

$$= 2 + \frac{3\theta}{2N} - \frac{4\theta}{2N} - \frac{3\alpha}{2}\left(1-\frac{\theta}{N}\right)$$

$$= 2 + \frac{3\theta-4\theta}{2N} - \frac{3\alpha}{2}\left(1-\frac{\theta}{N}\right)$$

$$\mu_2^{'} = 2 - \frac{\theta}{2N} - \frac{3\alpha}{2}\left(1-\frac{\theta}{N}\right) \to 2$$

The sample frequency distribution is:

x:	0	1	2
f:	27	38	10

N = 75

$$\mu_1^{'} = \frac{1}{N}\Sigma fx = \frac{1}{75}\left[38+20\right] = \frac{58}{75}$$

$$\mu_2^{'} = \frac{1}{N}\Sigma fx^2 = \frac{1}{75}\left[38+40\right] = \frac{78}{75}$$

Equation the sample moments to theoretical moments, we get

$$1-\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right) = \frac{58}{75}$$

$$\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right) = 1-\frac{58}{75}$$

$$\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right) = \frac{17}{75} \qquad \rightarrow 3$$

$Sub\ equ\ 3\ becomes\ 2$

$$2-\frac{\theta}{2N}-\frac{3}{2}\alpha\left(1-\frac{\theta}{N}\right) = \frac{78}{75}$$

$$2-\frac{\theta}{2N}-3\left(\frac{17}{75}\right) = \frac{78}{75}$$

$$2-\frac{\theta}{2N}-\frac{51}{75} = \frac{78}{75} \Rightarrow \frac{\theta}{2N} = 2-\frac{51}{75}-\frac{78}{75}$$

$$\frac{\theta}{2N} = 2-\frac{129}{75}$$

$$\frac{\theta}{2N} = \frac{150-129}{75} \Rightarrow \frac{\theta}{2N} = \frac{21}{75}$$

$$\theta = \frac{21}{75}\times 2N \Rightarrow \hat{\theta} = \frac{42}{75}.N$$

$Sub\ \hat{\theta}\ value\ in\ equ\ 3$

$$\frac{\alpha}{2}\left(1-\frac{42}{75N}.N\right) = \frac{17}{75}$$

$$\frac{\alpha}{2}\left(\frac{75-42}{75}\right) = \frac{17}{75}$$

$$\frac{\alpha}{2}\left(\frac{33}{75}\right) = \frac{17}{75}$$

$$\frac{\alpha}{2} = \frac{17}{75}\times\frac{75}{33} \Rightarrow \frac{\alpha}{2} = \frac{17}{33}$$

$$\alpha = \frac{17}{33}\times 2$$

$$\hat{\alpha} = \frac{34}{33}$$

PROBLEM:3

Estimate α and β in the case of pearson's type III distribution by the method of moments.

$$f(x;\alpha,\beta)=\frac{\beta^{\alpha}}{\overline{|\alpha}}x^{\alpha-1}e^{-\beta x},0\le x\le\infty$$

Solution:

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

$$= \int_0^\infty \frac{x^r \beta^\alpha x^{\alpha-1} e^{-\beta x}}{\lceil \alpha} dx$$

$$= \frac{\beta^\alpha}{\lceil \alpha} \int_0^\infty x^{\alpha+r-1} e^{-\beta x}$$

$$= \frac{\beta^\alpha}{\lceil \alpha} \cdot \frac{\lceil \alpha + r}{\beta^{\alpha+r}} \qquad \because \textit{property of Beta 2 function}$$

$$= \frac{\beta^\alpha}{\lceil \alpha} \cdot \frac{\lceil \alpha + r}{\beta^\alpha \beta^r}$$

$$\mu_r^{'} = \frac{\lceil \alpha + r}{\lceil \alpha . \beta^r} \qquad \because \lceil n+1 = n \lceil n$$

$put\ r = 1$

$$\mu_1^{'} = \frac{\lceil \alpha + 1}{\lceil \alpha . \beta^1} \Rightarrow \mu_1^{'} = \frac{\alpha \lceil \alpha}{\lceil \alpha . \beta}$$

$$\boxed{\mu_1^{'} = \frac{\alpha}{\beta}}$$

$put\ r = 2$

$$\mu_2^{'} = \frac{\lceil \alpha + 2}{\lceil \alpha . \beta^2} \qquad \because \lceil n+2 = n+1 \lceil n+1$$

$$\mu_2^{'} = \frac{(\alpha+1)\lceil \alpha+1}{\lceil \alpha . \beta^2} \Rightarrow \frac{(\alpha+1)\alpha \lceil \alpha}{\lceil \alpha . \beta^2}$$

$i)\ Estimates\ of\ \alpha.$

$$\frac{\mu_2^{'}}{\left(\mu_1^{'}\right)^2} = \frac{\dfrac{\alpha(\alpha+1)}{\beta^2}}{\left(\alpha/\beta\right)^2} = \frac{\alpha(\alpha+1)}{\beta^2} \cdot \frac{\beta^2}{\alpha^2} = \frac{\alpha+1}{\alpha}$$

$$\frac{\mu_2^{'}}{\left(\mu_1^{'}\right)^2} = \frac{\alpha}{\alpha} + \frac{1}{\alpha} \Rightarrow \left(1 + \frac{1}{\alpha}\right)$$

$$\frac{\mu_2^{'}}{\left(\mu_1^{'}\right)^2} - 1 = \frac{1}{\alpha} \Rightarrow \frac{\mu_2^{'} - \left(\mu_1^{'}\right)^2}{\left(\mu_1^{'}\right)^2} = \frac{1}{\alpha}$$

$$\int_0^\infty x^r e^{-ax} dx = \frac{\lceil r+1}{a^{r+1}}$$

$r = \alpha + r - 1$

$r + 1 = \alpha + r - 1 + 1$

$a = \beta$

$$\alpha = \frac{\left(\mu_1'\right)^2}{\mu_2' - \left(\mu_1'\right)^2}$$

$$\hat{\alpha} = \frac{\left(m_1'\right)^2}{m_2' - \left(m_1'\right)^2}$$

ii) Estimates of β :

$$\mu_1' = \frac{\alpha}{\beta}$$

$$\beta = \frac{\alpha}{\mu_1'} \Rightarrow \beta = \frac{1}{\mu_1'} \cdot \frac{\left(\mu_1'\right)^2}{\mu_2' - \left(\mu_1'\right)^2}$$

$$\beta = \frac{\left(\mu_1'\right)}{\mu_2' - \left(\mu_1'\right)^2}$$

$$\hat{\beta} = \frac{m_1'}{m_2' - \left(m_1'\right)^2}$$

Here m_1' *and* m_2' *are sample moments.*

INTERVAL ESTIMATION:

Interval estimate means the population parameter given by two numbers between which the parameter is considered. Generally, point estimation does not confidently lie down our information. Therefore, two values are computed in such a way that the interval lies between the two values containing the parameter. An interval so obtained is called interval estimate or confidence interval.

For instance, studying a sample, we estimate that the average salary of a factory worker is Rs. 600; it is a point estimate. At the same time we may estimate through a sample study that an average salary of factory workers can lie between Rs. 600 and Rs. 700; this is an interval estimate.

UNIT-III

UNIT-III

Test of Hypothesis:

A statistical test is a rule or method which leads to make a decision whether to accept to accept or reject the hypothesis on the basis of a sample from a population.

Statistical Hypothesis:

Any statement or assumption about a population from which give a random sample may have been drawn is called statistical hypothesis.

Simple Hypothesis:

If the statistical hypothesis specifies the population completely then it is termed as a simple statistical hypothesis.

Composite Hypothesis:

If the statistical hypothesis does not specifies the population completely then it is termed as a Composite statistical hypothesis.

Example: If $x_1, x_2, \ldots\ldots x_n$ is a random sample of size n from a normal population with mean μ and variance σ^2, then the hypothesis $H_1: \mu = \mu_0, \sigma^2 = {\sigma_0}^2$ is a simple hypothesis, where as each of the following hypothesis is a composite hypothesis.

i) $\mu = \mu_0$, ii) $\sigma^2 = {\sigma_0}^2$ iii) $\mu < \mu_0, \sigma^2 = {\sigma_0}^2$ iv) $\mu > \mu_0, \sigma^2 = {\sigma_0}^2$

v) $\mu = \mu_0, \sigma^2 < {\sigma_0}^2$, vi) $\mu = \mu_0, \sigma^2 > {\sigma_0}^2$ $vi) \mu < \mu_0, \sigma^2 > {\sigma_0}^2$

Null Hypothesis:

A definite statement about the population parameter is called the null hypothesis and it sis denoted by H_0. $H_0: \mu = \mu_0$

Alternative Hypothesis:

Any hypothesis which is complementary to the null hypothesis is called an alternative and is usually denoted by H_1. $i)$ $H_1 : \mu \neq \mu_0$ ii) $H_1 : \mu > \mu_0$ iii) $H_1 : \mu < \mu_0$

Critical Region: (Rejection region)

The region of rejection of the null hypothesis (H_0) when H_0 is true is called critical region.

Acceptance region:

The region of acceptance of the null hypothesis (H_0) when H_0 is false is called a acceptance region.

Meaning of Critical Region:

Let $x_1, x_2, \ldots\ldots x_n$ be the sample observation denoted by O. All the values of O will be aggregate of a sample and they constitute a space, called the sample space, which is denoted by S.

Since the sample values $x_1, x_2, \ldots\ldots x_n$ can be taken as a point in 'n' dimensional space, we specify some region of the 'n' dimensional space and see whether this points $x_1, x_2, \ldots\ldots x_n$

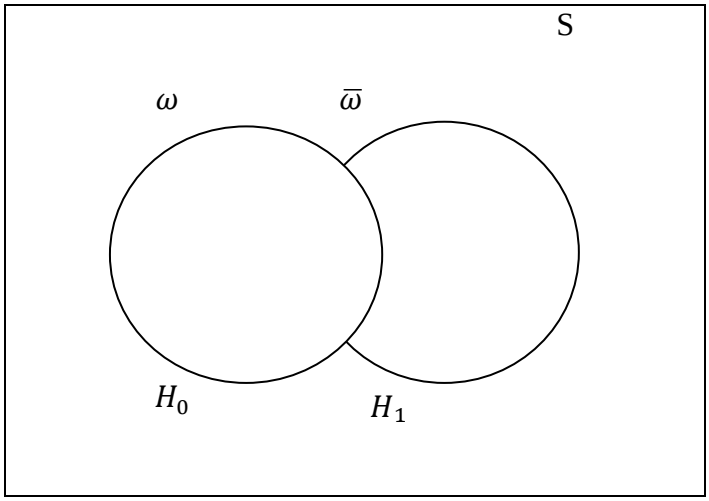
falls within this region (or) outside of this region. We divide the sample space 'S' into two disjoint parts say $\bar{\omega}$ and $\bar{\omega} = s - \omega$ (or) $\bar{\omega}$ or ω.

If the sample point falls in ω, we reject the null hypothesis H_0 and sample points fall in $\bar{\omega}$, We reject the hypothesis H_1 and accept the null hypothesis H_0. This is shown in the following diagram.

Critical region

(or)

Rejection region



Acceptance

The size of critical region is called the level of significance and it is denoted by 'α'.

Whether α is the probability of rejecting H_0. When H_0 is true symbolically.α = P(x$\in$ ω/ H_0)

Type –I error:

The error of rejecting H_0 $(accepting\ H_1)$ when H_0 is true is called Type- I error.

Type –II error:

The error of accepting H_0 when H_0 is false (H_1 is true) is called Type- II error.

The four possible situations that arise in a test of hypothesis are expressed in the following table:

Actual	Decision	
	Accept H_0	$Reject\ H_0$
H_0 is true	Correct decision(no error) Probability =1-α	Wrong(Type I error) Probability = α
H_0 is false	Wrong(Type II error) Probability =β	Correct decision (no error) Probability = 1- β

The probability of Type I and Type II errors are denoted by $\alpha\ and\ \beta$ respectively.

Thus α = P (type I error)

= P(Rejecting H_0 when H_0 is true)

β = P (type II error)

= P(Accepting H_0 when H_0 is false)

Symbolically, α = P (x $\epsilon\omega/H_0$) = $\int_{\omega} L_0$ dx

β = P (x $\epsilon\bar{\omega}/H_1$) = $\int_{\bar{\omega}} L_1$ dx

Where L_1 is the likelihood function of the sample observation under H_1.

Since $\int_{\omega} L_1$ dx + $\int_{\bar{\omega}} L_1$ dx =1

$\int_{\omega} L_1$ dx = 1- $\int_{\bar{\omega}} L_1$ dx = 1- β $\Rightarrow$= 1- β = $\int_{\omega} L_1$ dx $\Rightarrow$ 1- β = P (x $\epsilon\omega/H_1$)

This 1- β is called the power of a test.

Level of Significance: (α)

The probability of type I error, is known as the level of significance of the test. It is also

Called the size of the critical region.

Power of the test:

$1-\beta$ is called the power of the test. Thus the power of a test is defined as P (sample points falls in the critical region ω) when H_1 is true and H_0 is false.

Neyman Pearson Lemma [NPL] Theorem:

This Lemma provides the most powerful test of simple hypothesis against a simple alternative hypothesis.

Statement: Let K > 0, be a constant and ω be a critical region of size α such that

$$\omega = \{x \in s : \frac{f(x,\theta_1)}{f(x,\theta_0)} > k\}$$

$$\omega = \{x \in s : \frac{L_1}{L_0} > k\} \rightarrow A$$

and $\bar{\omega} = \{x \in s : \frac{L_1}{L_0} < k\} \rightarrow B$

Where L_0 and L_1 are the likelihood function of the sample observation x= ($x_1, x_2, \ldots\ldots x_n$) under $H_0 and\ H_1$ respectively.

Then ω is the most powerful critical region of the test hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$.

Proof: We are given that $\frac{L_1}{L_0} > k$

$$P(x \epsilon \omega / H_0) = \int_{\omega} L_0 \, dx = \alpha \rightarrow (1)$$

The power of the region is

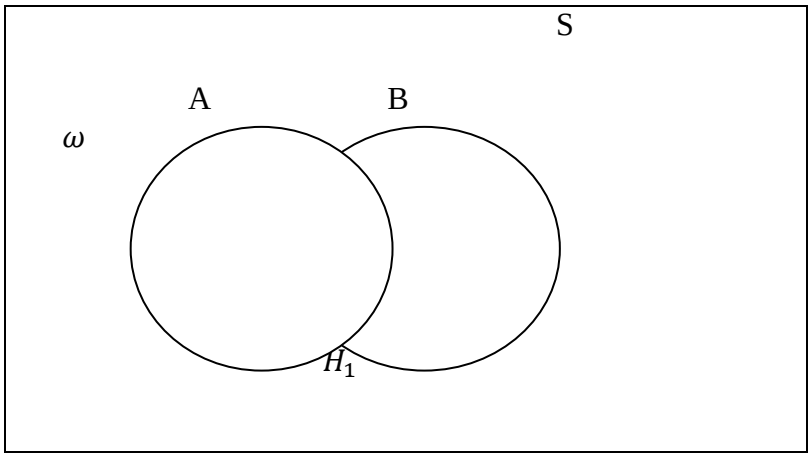
$$P(x \epsilon \omega / H_1) = \int_{\omega} L_1 \, dx = 1 - \beta \rightarrow (2)$$

In order to establish the lemma, we have to prove that there exists no other critical region of the less than or equal to α, Which is more powerful than ω. Let $\omega_1$be another critical region of size $\alpha_1 \leq \alpha$ and power 1- β_1 so that we have

$$P(x \epsilon \omega_1 / H_0) = \int_{\omega_1} L_0 \, dx = \alpha_1 \rightarrow (3)$$

And $P(x \epsilon \omega_1 / H_1) = \int_{\omega_1} L_1 \, dx = 1 - \beta_1 \rightarrow (4)$

Now we have to prove that $1 - \beta \geq 1 - \beta_1$ Consider the sample space.



From the figure, we have (C may be empty, (i.e) ω & ω_1 may be disjoint)

If $\alpha_1 \leq \alpha$, we have $\quad P(x \epsilon \omega_1 / H_0) \leq P(x \epsilon \omega / H_0)$

$$\int_{\omega_1} L_0 \, dx \leq \int_{\omega} L_0 \, dx$$

$$\int_{B \cup C} L_0 \, dx \leq \int_{A \cup C} L_0 \, dx$$

$$\int_B L_0 \, dx + \int_C L_0 \, dx \leq \int_A L_0 \, dx + \int_C L_0 \, dx$$

$$\int_B L_0 \, dx \leq \int_A L_0 \, dx \quad [\because \text{C is empty}]$$

$$\Rightarrow \int_A L_0 \, dx + \int_B L_0 \, dx \rightarrow (5)$$

Since A $\subset \omega$, (A)$\Rightarrow \frac{L_1}{L_0} > k \quad \Rightarrow L_1 > L_0 \, k$

$$\int_A L_1 \, dx > K \int_A L_0 \, dx$$

Multiply K in equation (5)

(5) x K $\Rightarrow$ K $\int_A L_1 \, dx \geq K \int_B L_0 \, dx \rightarrow$ (*)

$$\int_A L_1 \, dx \geq K \int_B L_0 \, dx \rightarrow (6) \quad [\text{using } (*)]$$

Also (B) $\Rightarrow \frac{L_1}{L_0} \leq k$

$$L_1 \leq K \, L_0 \quad \forall \quad x \epsilon \bar{\omega}$$

$$\int_{\bar{\omega}} L_1 \, dx \leq K \int_{\bar{\omega}} L_0 \, dx \rightarrow (7)$$

This result also holds for any subset of $\bar{\omega}$, since B $\subset \bar{\omega}$, we get

$$\int_B L_1 \, dx \leq K \int_B L_0 \, dx \rightarrow (8)$$

Equation (6) & (8) $\Rightarrow \int_B L_1 \, dx \leq \int_A L_1 \, dx$

Adding $\int_C L_1 \, dx$ on both sides we get. $\int_B L_1 \, dx + \int_C L_1 \, dx \leq \int_A L_1 \, dx + \int_C L_1 \, dx$

$$\Rightarrow \int_{B \cup C} L_1 \, dx \leq \int_{A \cup C} L_1 \, dx$$

$$\int_{\omega_1} L_1 \, dx \leq \int_{\omega} L_1 \, dx$$

$$P(x \in \frac{\omega_1}{H_1}) \leq P(x \in \frac{\omega}{H_1})$$

$$1 - \beta_1 \leq 1 - \beta$$

$$1 - \beta \geq 1 - \beta_1$$ Hence the lemma.

Problem: 1

Given the frequency function: $f(x, \theta) = \begin{cases} \frac{1}{\theta} & ; 0 \leq x \leq \theta \\ 0 \quad ; & elsewhere \end{cases}$

And that you are testing the null hypothesis $H_0 : \theta=1$ against $H_1 : \theta=2$, by means of a single observed value of x. What would be the sizes of the type I and type II errors, if you choose the interval (i) $0.5 \leq x$, (ii) $1 \leq x \leq 1.5$ as the critical regions? Also obtain the power function of the test.

Solution: Here we want to test $H_0 : \theta=1$, against $H_1 : \theta=2$

(i) Here $\omega = \{ x: 0.5 \leq x\} = \{ x: x \geq 0.5\}$ and $\bar{\omega} = \{ x: x \leq 0.5 \}$

$$\alpha = P(x \in \omega / H_0) = \int_{\omega} L_0 \, dx \, / H_0 = \int_{0.5}^{\theta} f(x, \theta) dx \, / H_0 : \theta=1$$

$$= \int_{0.5}^{1} \frac{1}{1} dx = \int_{0.5}^{1} dx = [x]_{0.5}^{1} = 1\text{-}0.5 = 0.5$$

$$\alpha = 0.5$$

Similarly, $\beta = P(x \in \bar{\omega} / H_1) = \int_{\bar{\omega}} L_1 \, dx / H_1 = \int_{0.5}^{\theta} f(x, \theta) dx \, / H_1 : \theta=2$

$$= \int_{0}^{0.5} \frac{1}{\theta} dx \, / H_1 : \theta=2 \Rightarrow \int_{0}^{0.5} \frac{1}{2} dx = 1/2 \; [x]_{0}^{0.5} = 0.5/2 = 0.25$$

$$\beta = 0.25$$

Thus the size of type I and type II errors are respectively $\alpha = 0.5$ and

$\beta = 0.25$ and power function of the test = 1- $\beta = 1 - 0.25 = 0.75$

ii) $\omega = \{ x: 1 \leq x \leq 1.5\}$

$$\alpha = P(x \in \omega / H_0) = \int_{\omega} L_0 \, dx \, / H_0 = \int_{1}^{1.5} f(x, \theta) dx \, / H_0 : \theta=1$$

$$= \int_{1}^{1.5} \frac{1}{\theta} dx \, / H_0 : \theta=1 = \int_{1}^{1.5} \frac{1}{1} dx = [x]_{1}^{1.5} = 1.5\text{-}1 = 0.5$$

$$\alpha = 0.5$$

Since under H_0 : θ=1, f(x, θ) = 0 , for $1 \leq x \leq 1.5$

β = P (x $\epsilon\bar{\omega}/H_1$)

We have P (x $\epsilon\omega/H_1$) + P (x $\epsilon\bar{\omega}/H_1$) = 1

P (x $\epsilon\bar{\omega}/H_1$) = 1 - P (x $\epsilon\omega/H_1$)

Consider, P (x $\epsilon\omega/H_1$) = $\int_{\bar{\omega}} L_1$ dx/H_1 = $\int_1^{1.5} f(x,\theta)dx$ /H_1: θ=2

$\Rightarrow \int_1^{1.5} \left(\frac{1}{\theta}\right) dx$ /H_1: θ=2 $\Rightarrow \int_1^{1.5} \frac{1}{2} dx = \frac{1}{2}[x]_1^{1.5}$ = 1/2 [1.5-1] = $\frac{0.5}{2}$ = 0.25

$\Rightarrow$ P (x $\epsilon\omega/H_1$) = 0.25

$\therefore$ β = P (x $\epsilon\bar{\omega}/H_1$) = 1 - P (x $\epsilon\omega/H_1$)

= 1 – 0.25 =0.7 $\Rightarrow$ β = 0.75

$\therefore$ The power function of the test = 1 – β

= 1 – 0.75 = 0.25

Problem: 2

If x $\geq$ 1 is the critical region for testing H_0: θ=2 against the alternative

H_1: θ=1 On the basis of the single observation from the population,

f(x, θ) = $\theta\ e^{-\theta x}$; $0 \leq x \leq \infty$,

Obtain the value of type I and type II errors.

Solution: H_0: θ=2 Vs H_1: θ=1

Here Given that ω = { x: x $\geq$ 1}

$\bar{\omega}$ = { x: x $\leq$ 1}

α = P (x $\epsilon\omega/H_0$) = $\int_{\bar{\omega}} L_0$ dx = $\int_1^{\infty} f(x,\theta)dx$ /H_0: θ=2

= $\int_1^{\infty} \theta\ e^{-\theta x} dx$ /H_0: θ=2 = $\int_1^{\infty} 2e^{-2x}\ dx$ = 2[$\frac{e^{-2x}}{-2}$]$_1^{\infty}$

= - [$e^{-\infty}$ – e^{-2}] = - [0 - e^{-2}] = e^{-2} = $\frac{1}{e^2}$

α = $\frac{1}{e^2}$

β = P (x $\epsilon\bar{\omega}/H_1$) = $\frac{\int}{\bar{\omega}} L_1$ dx = $\int_0^1 f(x,\theta)dx$ /H_1: θ=1

= $\int_0^1 \theta\ e^{-\theta x}dx$ /H_1: θ=1 $\Rightarrow$ $\int_0^1 1.e^{-x}dx$

= $[\frac{e^{-x}}{-1}]_0^1$ = - [e^{-1} – e^0] = - [1/e -1] = - 1/e + 1

= 1 – 1/e = $\frac{e-1}{e}$

Problem: 3

Let P be the probability that a coin will fall head in a single toss in order to test H_0 = P = 1/2 and H_1 = P = 3/4. The coin is tossed 5 times and H_0 is rejected if more than 3 heads are obtained. Find the probability of type I error and power of the test.

Solution: H_0 = P = 1/2 and H_1 = P = 3/4.

If the random variable X denotes the no. of. Heads in n tosses of a coin then

X ~B(n, p) so that

P(X= x) = $nc_x\ p^x q^{n-x}$

= $nc_x\ p^x(1-p)^{n-x}$

P(x) = $5c_x\ p^x(1-p)^{5-x}$

The critical region is given by : ω = { x: x $\geq$ 4}

$\bar{\omega}$ = { x: x $\leq$ 3}

α = P (x $\epsilon\omega/H_0$) = $\frac{\int}{\omega} L_0$ dx/H_0 : p = 1/2

= $\sum_{X=4}^{5} p(x)/\ H_0$: p = 1/2

= p(x=4) + p(x=5)/ H_0 : p = 1/2

= $5C_4\ (1/2)^4\ (1-1/2)^{5-4}$ + $5C_5\ (1/2)^5\ (1-1/2)^{5-5}$

= 5 $(1/2)^4\ (1/2)^1$ + 1 $(1/2)^5\ (1/2)^0$

= 5$(1/2)^5$ + $1/2)^5$ = 6 $(1/2)^5$ = 3/16

α = 0.1875

β = P (x $\epsilon\bar{\omega}/H_1$) = 1 – P (x $\epsilon\omega/H_1$)

= 1 –[$\sum_{X=4}^{5} p(x)/\ H_1$: p = 3/4]

$= 1 - [\{5C_4\ (3/4)^4\ (1-3/4)^{5-4}\ + 5C_5\ (3/4)^5\ \ (1-3/4)^{5-5}\ ]$

$= 1 - \left[\{5C_4\ (3/4)^4\ (\frac{1}{4})\ + 5C_5\ (3/4)^5\ \}\right]$

$= 1 - (3/4)^4\ \left\{\frac{5}{4} + \frac{3}{4}\right\}$

$= 1 - (3/4)^4\ \left[\frac{8}{4}\right]$

$= 1 - \frac{81}{128}\ =\ \frac{47}{128}\ = 0.3672$

$\beta =\ 0.3672$

Power of the test = $1 - \beta$

$= 1\ - 0.3672$

$= 0.6328$

Problem: 4

Let X has a p. d. f of the form:

$$f(x, \theta) = \begin{cases} \frac{1}{\theta}\ e^{-\theta x}; 0 < x < \infty, \theta > 0 \\ 0\ ,elsewhere \end{cases}$$

$H_0 : \theta\ = 2$, against $H_1 : \theta\ = 1$, use the random sample x_1, x_2 of the size 2 and define a critical region: $\omega = \{(x_1,\ x_2) : 9.5 \leq x_1 + x_2\}$.

Find: (i) Power of the test

(ii) Significance level of the test.

Solution: We are given the critical region:

$\omega = \{(x_1,\ x_2)\colon 9.5 \leq x_1 + x_2\} = \{(x_1,\ x_2)\colon\ x_1 + x_2 \geq 9.5\}$

Size of the critical region (i.e) the significance level of the test is given by:

$\alpha =$ P (x $\epsilon\omega / H_0$) = $P\ [\ x_1 + x_2 \geq 9.5 / H_0] \rightarrow (*)$

In sampling from the given exponential distribution,

$\frac{2}{\theta}\sum_{i=1}^{n} x_i \sim \chi_{(2n)}^{\ 2}$ $\Rightarrow$ U = $\frac{2}{\theta}(x_1 + x_2\ ) \sim \chi_{(4)}^{\ 2}$, (n=2)

$\therefore\ \alpha$ = P$\left[\frac{2}{\theta}(x_1 +\ x_2\ ) \geq \frac{2}{\theta} X 9.5 / H_0\right]$ $\rightarrow$ [From (*)

$= P\left[\chi_{(4)}^{2} \geq 9.5 \right]$

$\alpha = 0.05$

Power of the test is given by

$1 - \beta = P(x \in \omega / H_1) = P\left[x_1 + x_2 \geq 9.5 / H_1 \right]$

$= P\left[\frac{2}{\theta} (x_1 + x_2) \geq \frac{2}{\theta} X 9.5 / H_1 \right]$

$= P\left[\chi_{(4)}^{2} \geq 19 \right]$

UNIT-IV

UNIT-IV

Test of Significance:

A very important aspect of the sampling theory is the study of tests of significance which enable us to decide on the basis of the sample results if,

- The deviation between the observed sample statistics and the hypothetical parameter value is significant.
- The deviation between the two sample statistics is significant.

Since for large n, almost all the distribution, Example: Binomial, Poisson, t, F, chi-square

can be approximated very closely by a normal probability curve, we use the Normal test of significance for large samples. Some of the well-known test of significance for studying such differences for small samples is t-test, F-test.

Sampling Distribution

The distribution of all possible values which can be assumed by some statistic measured from samples of same size 'n' randomly drawn from the same population of size N is called as sampling distribution of the statistic (DANIEL and FERREL).

Consider a population with N values. Let us take a random sample of size n from this population, and then there are

$NC_n = \frac{N!}{n!(N-n)!} = k$ (Say), possible samples. From each of these k samples if we compute a statistic (e.g. mean, variance, correlation coefficient, skewness etc) and then we form a frequency distribution for these k values of a statistic. Such a distribution is called sampling distribution of that statistic.

For example, we can compute some statistic $t = t(x_1,x_2,\ldots,x_n)$ for each of these k samples. Then $t_1,t_2,\ldots t_k$ determine the sampling distribution of the statistic t. In other words statistic t may be regarded as a random variable which can take the values $t_1,t_2,\ldots t_k$ and we can compute various statistical constants like mean, variance, skewness, kurtosis etc., for this sampling distribution.

Standard Error:

The standard deviation of the sampling distribution of a statistic is known as its standard error.

S. No	Statistics	Standard Error (S.E)
1.	Sample Mean ($\bar{x}$)	$\frac{\sigma}{\sqrt{n}}$
2.	Sample Proportion (P)	$\sqrt{PQ/n}$
3.	Sample standard deviation (S)	$\sqrt{\sigma^2/2n}$
4.	Sample variance (S^2)	$\sigma^2\sqrt{2/n}$
5.	Difference of two sample mean=($\bar{x}_1$-$\bar{x}_2$)	$\sqrt{\frac{\sigma^2{}_1}{n_1}+\frac{\sigma^2{}_2}{n_2}}$

Utility of Standard Error:

1. To determine the precision of the sample estimate of some population parameter, which is given y the reciprocal of the S.E of the sampling distribution of the estimate .

   Thus, if t is a statistic used to estimate the parameter θ then Precision of t= 1/S.E (t).
2. To test the significance of the difference between two independent sample estimates of the same population parameter.
3. To Obtain point estimates of the population parameter.
4. To obtain interval estimates of the population parameter.

   (i.e) to obtain probable limits between which the true value of the parameter may be expected to lie.

Standard error:

The Standard deviation of the sampling distribution of a statistic is known as its standard error. It is abbreviated as S.E. For example, the standard deviation of the sampling distribution of the mean $\bar{x}$ known as the standard error of the mean,

Where $v(\bar{x}) = v\left(\frac{x_1+x_2+\cdots+x_n}{n}\right)$

$$= \frac{v(x_1)}{n^2} + \frac{v(x_2)}{n^2} + \cdots + \frac{v(x_n)}{n^2}$$

$$= \frac{\sigma^2}{n^2} + \frac{\sigma^2}{n^2} + \cdots + \frac{\sigma^2}{n^2} = \frac{n\sigma^2}{n^2}$$

$\therefore$ The S.E of the mean is $\frac{\sigma}{\sqrt{n}}$

Uses of standard error:

i) Standard error plays a very important role in the large sample theory and forms the basis of the testing of hypothesis.
ii) The magnitude of the S.E gives an index of the precision of the estimate of the parameter.
iii) The reciprocal of the S.E is taken as the measure of reliability or precision of the sample.
iv) S.E enables us to determine the probable limits within which the population parameter may be expected to lie.

One tail and two tailed test:

Two tail test is one where the hypothesis about the population parameter is rejected for the value of sample statistic falling into the either tails of the sampling distribution.

One tailed test:

When the hypothesis about the population parameter is rejected only for the value of sample statistic falling into one of the tails of the sampling distribution. Then it is known as one – tailed test.

Two tailed test:

A test of statistical hypothesis where the alternative hypothesis is two tailed such as, H_0: $\mu = \mu_0$ against the alternative hypothesis H_1: $\mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$) is known as two tailed test and in such a case the critical region is given by the proportion of the area lying in both the tails of the probability curve of test of statistic.

Example for one – tailed and two tail test:

Let us suppose that there are two popular brands of bulbs one manufactured by standard process (with mean life μ_1) and the other manufactured by new process (with mean life μ_2).

If our test is whether the bulbs differ significantly then our hypothesis is $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$. This gives two tail.

If the bulb produced by the new process have a higher average life than those produced by standard process. $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 < \mu_2$. This gives left tail.

If the new process is inferior to that of standard process we have $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$. This gives right – tail test.

Errors - type I and type II error:

When the hypothesis is tested, there are four possible results.

- The hypothesis is true but our test reject it
- The hypothesis is false but our test accepts it
- The hypothesis is true our test accepts it
- The hypothesis is false but our rejects it

Condition		
Decision	H_0 : True	H_0: false
Accept H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Type I Error:

Reject H_0 when it is true

P{Type I error} = $\propto$ i.e. P{Reject a lot when it is good} = $\propto$. Producer's risk

Type II Error:

Accept H_0 when H_1 is true.

P{Type II error} = β P{Accept a lot when it is bad} = β. Consumer's risk

Level of significance:

The statistical test fix the probability of committing type I error at a certain called the level of significance and minimize the chances of committing the type II error.

Critical region and level of significance:

- Critical region is a region of rejection of H_0.
- Level of significance is the probability $\propto$ that a random value of the test statistic belongs to the critical region is known as level of significance.
- The level of significance usually employed in testing of hypothesis is 5% (or) 1%. The level of significance is always fixed in advance before collecting the sample information.

Critical values (or) significant values:

The value of statistics which divides the critical region (or) rejection region and the acceptance region is called the critical value (or) significant value. It depends on.

1. Level of significance
2. Alternative hypothesis

Critical value

Type of test	Level of significance			
	1%	**2%**	**5%**	**10%**
Two – tailed test	$\|z_\propto\|$ = 2.58	2.33	1.96	1.645
Right – tailed test	2.33	2.055	1.645	1.28
Left – tailed test	-2.33	-2.055	-1.645	-1.28

For large samples the standardized variable corresponding to the statistic

$$z = \frac{t - E(t)}{S.E(t)} \sim N(0,1) as\ n \to \infty.$$

Procedure for testing the hypothesis:

Step 1: Null hypothesis: Set up the Null hypothesis H_0.

Step 2: Alternative hypothesis: Set up the alternative hypothesis. H_1.

Step 3: Level of significance: Set up a suitable level of significance 5% or 1%

Choose the appropriate level of significance ($\propto$) depending on the reliability of the estimates and permissible risk. This is to be decided before sample is drawn. (i.e) $\propto$ is fixed in advance.

Step 4: Test statistic (or test criterion) Determine a suitable test statistic

$$z = \frac{t - E(t)}{S.E\ (t)}\text{ , under } H_0$$

Step 5: Conclusion (or) Inference. We compare the computed value of Z with the significant value (tabulated value) Z_α at the given level of significance, 'α'.

If $|z_\propto| < Z_\alpha$, accepted H_0.

If $|z_\propto| > Z_\alpha$, rejected H_0.

LARGE SAMPLE TEST:

When the sample size is equal to or greater than 30, ($n \geq 30$) then the sample is called large sample test.

We consider the following tests under large sample test.

- Z-test for single mean.
- Z-test for difference of two means.
- Z-test for single proportion.
- Z-test for difference of two proportion.

Z-test for single mean:

Let x_i(i = 1, 2,…n) be a random sample of size n from a population with variance σ^2, then the sample mean $\bar{x}$ is given by

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

$$E(\bar{x}) = \mu$$

$$V(\bar{x}) = V\left[\frac{1}{n}(x_1 + x_2 + \cdots + x_n)\right]$$

$$= \frac{1}{n^2}[V(x_1) + V(x_2) + \cdots + V(x_n)]$$

$$= \frac{1}{n^2} n\sigma^2 \qquad = \frac{\sigma^2}{n}$$

$$\therefore \quad \text{S.E}(\bar{x}) \qquad = \frac{\sigma}{\sqrt{n}}$$

Test of Significance for single mean:

If x_i , i = 1, 2, 3 ….n is a random sample of size 'n' from a normal population with mean μ and variance σ^2. Suppose we want to test whether the samples have been drawn from the population with mean μ and variance σ^2.

The null hypothesis (H_0): The sample has been drawn from a population with mean μ and variance σ^2.

i.e) $H_0 : \mu = \mu_0$

Alternative hypothesis (H_1): The sample has not been drawn from a population with mean μ and varianceσ^2.

i.e) $H_1 : \mu \neq \mu_0$ $(\mu > \mu_0 0r \mu < \mu_0)$

Test statistics: under H_0, the test statistic is

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Where $\bar{x}$ - sample mean, μ - population mean, σ - population S.D , n- sample size.

Level of significance: The L.o.s (α) which indicates whether the probability of difference is small or large is generally fixed. (5% or 1%)

Conclusion:

If the computed value of $|z|$ is less than the critical value of Z, i.e) $|z| < z_\alpha$, we accept our H_0.

If the computed value of $|z|$ is greater than the critical value of Z, i.e) $|z| > z_\alpha$, we reject our H_0.

Problem: 1

A random sample of 400 male students is found to have a mean height of 171.38cms. Can it be reasonable regarded as a sample from large population with mean height 171.17cms and S.D 3.30cms?

Solution:

Sample size n = 400

Sample mean $\bar{x}$ = 171.38 cm

Population mean μ = 171.17cm

Population S.D σ = 3.30 cm

Step 1:

Null hypothesis (H_0): The sample has been drawn from the normal population with mean

μ = 10.2 cm. and S.D σ = 3.30 cm

i.e, $H_0 : \mu = 10.2$

Step 2:

Alternative hypothesis H_1: $\mu \neq 10.2$ (two-tailed test)

Step 3:

Under H_0, The test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{171.38 - 171.17}{3.30/\sqrt{400}} = 1.27$$

Step 4:

Table value: The table value of Z at 5% L.O.S = 1.96

Step 5:

Conclusion: Since $|z| = 1.7777 < z_\alpha = 1.96$ .Since calculated value is less than the table value H_0 is accepted. Otherwise reject it.

∴ We can conclude that the sample has been drawn from the normal population with mean

$\mu = 10.2$ cm and S.D $\sigma = 3.30$ cm.

Problem: 2

The mean breaking strength of the cables supplied by a manufacturer is 1800 with an SD of 100. By a new technique in the manufacturing process, it is claimed that the breaking strength of the cable has increased. To test this claim a sample of 50 cables is tested and is found that the mean breaking strength is 1850. Can we support the claim at 1% level of significance?

Solution:

Given n = 50, $\bar{x} = 1850$, $\mu = 1800$, $\sigma = 100$

Step 1:

Null hypothesis (H_0): the mean breaking strength of the cables is 1800. The sample mean do not differ significantly.

i.e) $H_0 : \mu = 1800$

Step 2:

Alternative hypothesis (H_1): The sample mean differ significantly.

i.e) H_1: $\mu > 1800$ (Right – tailed test)

Step 3:

Test statistic: Under H_0, the test statistic is

$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1850-1800}{\frac{100}{\sqrt{50}}} = \frac{50}{14.1421} = 3.54$$

Step 4:

Level of significance: The table (critical) value of Z at $\alpha = 1\%$ level of significance is

$z_\alpha = 2.33$.

Step 5:

Conclusion:

Since $|z| > z_\alpha$, we reject the H_0 and accept H_1, i.e. the mean breaking strength of the cable has increased.

Problem: 3

A sample of 100 items, drawn from a universe with mean value 64 and S.D. 3, has a mean value 63.5. Is the difference in the means sidnificant?

Solution:

Given n = 100, $\bar{x} = 63.5$, , $\mu = 64$, $\sigma = 100$

Step 1:

Null hypothesis (H_0): The sample mean do not differ significant.

i.e) H_0 : μ =64

Step 2:

Alternative hypothesis (H_1): The sample mean is differ significant.

i.e) H_1: $\mu \neq 64$

Step 3:

Test statistic: Under H_0, the test statistic is

$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{63.5-64}{3/\sqrt{100}} = \frac{-0.5}{0.3} = -1.6667$$

$$|z| = 1.6667$$

Step 4: Level of significance:

The table (critical) value of Z at $\alpha = 5\%$ level of significance is $z_\alpha = 1.96$

Step 5:

Conclusion:

$|z| = 1.6667 < z_\alpha = 1.96$

Since the calculated value is less than table value, we accept our H_0. Otherwise reject it,

∴ We can conclude that the sample mean do not differ significant.

Test of Significance for difference between two means:

Test procedure: Let $\bar{x}_1$ be the mean of a sample of size n_1 from a population with mean μ_1 and variance ${\sigma_1}^2$ 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${\sigma_2}^2$.

Suppose we want to test whether the two samples have been drawn from the same population we should use the test of significance.

Step 1:

The null hypothesis (H_0): There is no significant difference between the sample means.
i.e) $H_0 : \mu_1 = \mu_2$

Step 2:

Alternative hypothesis (H_1): There is significant difference between the sample means.

i.e) $H_1: \mu_1 \neq \mu_2 (\mu_1 > \mu_2 0r \mu_1 < \mu_2)$

Step 3:

The test statistic

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Step 4:

The Level of significance (α) which indicates whether the probability of difference is small or large is generally fixed. (α = 5% or 1%)

Step 5:

Conclusion:

If the computed value of $|z|$is less than the critical value of z,

i.e) If $|z| < z_\alpha$, accepted H_0 .

If the computed value of $|z|$is greater than the critical value of z,

i.e) If $|z| > z_\alpha$, rejected H_0 .

Problem: 1

A buyer of electric bulbs bought 100 bulbs each of two famous brands. Upon testing these he found that brand A had a mean life of 1500 hours with a standard deviation of 50 hours whereas brand B had a mean life of 1530 hours with a standard deviation of 60 hours. Can it be concluded at 5% level of significance, that the two brands differ significantly in quality?

Solution:

We are given $\bar{x}_1 = 1500, \bar{x}_2 = 1530$, $s_1 = 50$, $s_2 = 60$, $n_1 = 100$ and $n_2 = 100$

Step 1:

Null hypothesis(H_0): The two brands of bulbs do not differ significantly in quality.

i.e) $H_0 : \mu_1 = \mu_2$

Step 2:

Alternative hypothesis(H_1) : the two brands of bulbs differ significantly.

i.e., H_1: $\mu_1 \neq \mu_2$ (two – sided alternative)

Step 3:

Under H_0, the Test statistic is,

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

$$= \frac{1500-1530}{\sqrt{\frac{50^2}{100}+\frac{60^2}{100}}} = -\frac{30}{7.81} = -3.84$$

Step 4:

The table value (critical value) of z at 5% LOS is $z_\alpha = 1.96$.

Step 5:

Conclusion: Since $|z| > z_\alpha$, the null hypothesis is rejected and hence we may conclude that the two brands of bulbs differ significantly in quality.

Problem: 2

The average hourly wage of a sample of 150 workers in plant 'A' was Rs. 2.56 with a standard deviation of Rs. 1.08. The average hourly wage of a sample of 200 workers in plant 'B' was Rs. 2.87 with a standard deviation of Rs.1.28. Can an applicant safely assume that the hourly wages paid by plant 'B' are higher than those paid by plant 'A'?

Solution:

In usual notations we are given

$n_1 = 150$, $n_2 = 150$, $\bar{x}_1 = 2.56$, $\bar{x}_1 = 2.87$, $s_1 = 1.08$ $s_2 = 1.28$

Step 1:

Null hypothesis (H_0): there is no significant difference between the mean level of wages of workers in plant A and plant B.

$$i.e) H_0 : \mu_1 = \mu_2$$

Step 2:

Alternative hypothesis (H_1): $\mu_1 < \mu_2$ (left – tailed test)

Step 3:

Test statistic: under H_0 the test statistic is

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

$$\therefore z = \frac{2.56 - 2.87}{\sqrt{\frac{(1.08)^2}{150} + \frac{(1.28)^2}{200}}} = -2.453$$

Step 4: Table value

The critical value of z for a one – tailed (left) test at 5% LOS is $z_\alpha = 1.645$

Conclusion:

Since $|z|$ = 2.453 is greater than the critical value $z_\alpha(1.645)$, the null hypothesis is rejected. Hence, we conclude that the average hourly wages paid by plant 'B' are certainly higher than those paid by Plant 'A'.

Problem: 3

The mean of two large samples of sizes 1000 and 2000 are 67.5 and 68.0 respectively. Test the equality of means of the two populations each with S.D 2.5. Assumptions should be stated clearly.

Solution:

Here $n_1 = 1000$, $n_2 = 2000$, $\bar{x}_1 = 67.5$, $\bar{x}_2 = 68$, and $\sigma_{1=}\sigma_2 = \sigma = 2.5$.

Step 1:

Null hypothesis (H_0): The samples have been drawn from the same population of s.d 2.5.

i.e) $H_0 : \mu_1 = \mu_2$

Step 2:

Alternative hypothesis (H_1): $\mu_1 \neq \mu_2$ (Two – tailed test)

Step 3:

Test statistic: under H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$\therefore Z = \frac{67.5 - 68}{2.5\sqrt{\frac{1}{1000} + \frac{1}{2000}}}$$

$$= \frac{-0.5}{2.5\sqrt{\frac{3}{2000}}} \quad = \frac{-0.5}{2.5\ X\ 0.039} = -5.13$$

$|z| = 5.13$

Step 4:

Table value: The critical value of z at $\alpha = 0.05$ LOS is $z_\alpha = 1.96$

Conclusion:

$Z = 5.13 > Z_\alpha = 1.96$

Since Calculated value is greater than the table value, we reject our H_0.Otherwise accept it.

∴ We can conclude that Samples are certainly not from the same population with s.d. 2.5.

Test of Significance for single proportion:

If X is the number of Success in n independent trials with Constant probability P of success for each trial. Let p be the proportion of the success and is given by p = $\frac{X}{n}$.

Suppose we want to test whether the sample proportion is significant or not, we have to set up,

Null hypothesis H_0: There is no significant difference between the sample proportion and the population proportion. i.e) H_0 **:** p = P_0

Alternative hypothesis H_1: There is significant difference between the sample proportion and the population proportion. i.e) p ≠ P_0 (P > P_0 or P < P_0)

Level of significance α = 5% or 1%

Test statistic:

Under H_0 , the test statistic is,

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$$

Where p – sample proportion, P- population proportion, Q = 1-P, n- sample size.

Level of significance: The LOS(α) which indicates whether the probability of difference is small or large is generally fixed.(α = 5% or 1%)

Conclusion:

If the computed value of $|z|$is less than the critical value of z,

i.e) If $|z| < z_\alpha$, accepted H_0 .

If the computed value of $|z|$is greater than the critical value of z,

i.e) If $|z| > z_\alpha$, rejected H_0 .

Problem:1

A manufacturer of light bulbs claims that an average 2% of the bulbs manufactured by his firm are defective. A random sample of 400 bulbs contained 13 defective bulbs. On the basis of this sample, can you support the manufacturer's claim at 5% level of significance?

Solution:

n = 400

X = 13

P = Sample proportion of defectives = $\frac{13}{400} = 0.0325$

Step 1:

Null hypothesis

H_0: = P = 0.02, i.e. 2% of bulbs are defective.

Step 2:

Alternative hypothesis

H_1: P > 0.02(Right – tailed test)

Step 3:

Level of significance $\alpha = 5\%$

Step 3:

Test statistic

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

$$= \frac{0.0325 - 0.02}{\sqrt{\frac{0.02 \times 0.98}{400}}}$$

z = 1.786

Step 4: Level of significance $\alpha = 5\%$

Critical value of z at 5% LOS for one tailed test = z_α= 1.645

Conclusion:

Since $|z| > z_\alpha$, we reject the null hypothesis H_0 and accept the alternative hypothesis H_1, i.e. there is a significant difference between the sample proportion and the claimed proportion i.e. the manufacturer's claim cannot be supported.

Problem: 2

In a sample of 1000 people in Mumbai, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance?

Solution:

We are given n = 1000

Let X = number of rice eaters = 540

∴ The sample proportion P = $\frac{x}{n} = \frac{540}{1000} = 0.54$

Step 1:

Null hypothesis

H_0: P = 0.5, i.e. both rice and wheat are equally popular in the state.

Step 2:

Alternative hypothesis

H_1: P ≠ 0.5(two – tailed alternative), i.e. Rice and wheat are not equally popular in the state.

Step 3:

The test statistic is $z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$

$$z = \frac{0.54 - 0.50}{\sqrt{\frac{0.5 \times 0.5}{1000}}} = 2.529$$

Step 4:

The significant or critical value of z at 1% level of significance for two tailed test is z_α= 2.58

Step 5:

Conclusion:

Since $|z| < z_\alpha$, We accept our null hypothesis and we may conclude that both rice and wheat are equally popular in this state.

Problem: 3

Twenty people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease, is 85% in favour of the hypothesis that it is more, at 5% level?

Solution:

Here n = 20, X= No. of. Persons who survived after attack by a disease = 18

p= Proportion of persons survived in the sample = $\frac{x}{n} = \frac{18}{20} = 0.90$

Step 1:

Null hypothesis (H_0) the proportion of persons survived after attack by a disease in the lot is 85% i.e) $H_0 : P = 0.85$

Step 2:

Alternative hypothesis (H_1): H_1: $P > 0.85$(Right – tailed alternative)

Step 3:

The test statistic is

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$$

Q= 1 – P = 1- 0.85 = 0.15

$$z = \frac{0.90 - 0.85}{\sqrt{\frac{0.85 \times 0.15}{20}}}$$

$$= \frac{0.05}{0.079} = 0.633$$

Step 4:

The significant or critical value of z at 5% level of significance for right tailed test is

z_α= 0.05 LOS is 1.645.

Step 5:

Conclusion: $Z = 0.633 < Z_\alpha = 1.645$

Since calculated value is less than table value we accept our H_0.

∴We can conclude that the proportion of persons survived after attack by a disease in the lot is 85%.

Test of Significance for difference of two proportion:

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. Then sample proportions are given by $p_1 = \frac{X_1}{n_1}$, $p_2 = \frac{X_2}{n_2}$.

Suppose we want to test the significance of the difference between the two proportions. We have to set up

Null hypothesis (H_0**):** There is no significant difference between the sample proportions.

i.e) H_0 : $p_1 = p_2$

Alternative hypothesis (H_1) : There is significant difference between the sample proportions.
i.e) H_1 : $p_1 \neq p_2$ ($p_1 > p_2$ or $p_1 < p_2$)

Test statistics:

Under the test statistic,

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \text{ ; } Q = 1 - p$$

Level of Significance : The l.o.s (α) which indicates whether the probability of difference is small or large is generally fixed.

Conclusion:

If the Computed value of $|Z|$ is less than the critical value of Z ,

i.e) If $z < z_\alpha$. H_0 is accepted, otherwise H_0 is rejected.

Problem: 1

A machine puts out 16 imperfect articles in a sample of 500. After the machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved?

Solution:

We are given $n_1 = 500$, $n_2 = 100$

p_1= Proportion imperfect articles before service = $\frac{16}{500} = 0.032$

p_2 = Proportion imperfect articles after service = $\frac{3}{100} = 0.03$

Step 1:

Null hypothesis

H_0: $P_1 = P_2$, i.e., the machine has not improved.

Step 2:

Alternative hypothesis

H_1: $P_1 > P_2$(Right – tailed test) i.e., there is significant improvement in the machine after overhauling.

Step 3:

Level of significance: $\alpha = 5\%$(say)

Step 4:

Test statistic

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Since P is not given, we estimate it as

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(500)(0.032) + (100)(0.03)}{600}$$

= 0.032 and Q = 1 – 0.032 = 0.968

$$z = \frac{0.032 - 0.03}{\sqrt{(0.032)(0.968)\left(\frac{1}{500} + \frac{1}{100}\right)}} = 0.1037$$

The critical value of z at 5% level of significance is $z_\alpha = 1.645$

Conclusion:

Since, $|z| \leq z_\alpha$ We accept our null hypothesis and hence we may conclude that the machine has not improved after overhauling.

Problem:2

In two large populations, there are 30 and 25 percent respectively of blue-eyed people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

Solution:

Here, we are given $n_1 = 1200$, $n_2 = 900$ and $P_1 = 30\% = 0.30$ and $P_2 = 25\% = 0.25$ $Q_1 = 0.70$ and $Q_2 = 0.75$

Step 1:

Null hypothesis H_0: $P_1 = P_2$, i.e., the difference in population proportions is likely to be hidden in sampling.

Step 2:

Alternative hypothesis H_1: $P_1 \neq P_2$ (two tailed alternative)

Step 3:

Level of significance: $\alpha = 5\%$ (say)

Step 4:

Test statistic

Under H_0: $P_1 = P_2$,

$$z = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} \sim N(0,1)$$

$$Z = \frac{0.30 - 0.25}{\sqrt{\frac{0.3 \times 0.7}{1200} + \frac{0.25 \times 0.75}{900}}} = 2.5538$$

The critical value or significant value at 5% LOS is $z_\alpha = 1.96$

Conclusion:

Since $|z| > 1.96$, i.e., 2.5538 > 1.96, we reject our null hypothesis at 5% LOS and we conclude that the difference in population proportions is unlikely to be hidden in sampling.

Problem:3

In a random sample of 100 men taken from village. A, 60 were found to be consuming alcohol. In another sample of 200 men taken from village B, 100 were found to be consuming

alcohol. Do the two villages differ significantly in respect of the proportion of men who consume alcohol?

Solution:

Given $n_1 = 100$, $n_2 = 200$

p_1= Sample proportion of men consuming alcohol in village A = $\frac{60}{100} = 0.6$

p_2 = Sample proportion of men consuming alcohol in village B = $\frac{100}{200} = 0.5$

Step 1:

Null hypothesis H_0: $P_1 = P_2$

Step 2:

The alternative hypothesis H_1: $p_1 \neq p_2$ (two – tailed test)

Step 3:

Level of significance: $\alpha = 5\%$ (say)

Step 4:

The test statistic is

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Where the estimate of P is

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(100)(0.60) + (200)(0.5)}{100 + 200}$$

$$= \frac{160}{300} = 0.533$$

$$\hat{Q} = 1 - \hat{P} = 0.467$$

$$\therefore z = \frac{0.6 - 0.5}{\sqrt{(0.533)(0.467)\left(\frac{1}{100} + \frac{1}{200}\right)}}$$

$$z = 1.6366$$

Step 5:

Conclusion:

Since $|z| < 1.96$, we accept the null hypothesis and conclude that the two villagers do not differ significantly in respect of the proportion of men who consume alcohol.

Problem: 4

A machine produced 20 defective articles in a batch of 400. After overhauling, it produced 10 defectives in a batch of 300. Has the machine improved?(Take α = 0.01)

Solution:

Here n_1 = 400, n_2 = 300, $x_1 = 20, x_2$ = 10.

Step 1:

Null hypothesis: There is no significant difference in the improvement of the machine before and after overhaul.

i.e) H_0: $p_1 = p_2$,

Step 2:

Alternative hypothesis: There is a significant difference in the improvement of the machine before and after overhaul.

i.e) H_1: $p_1 > p_2$ (Right – tailed test)

Step 3:

Test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

p_1= Sample proportion of defective articles before overhaul = $\frac{20}{400} = 0.05$

p_2 = Sample proportion of defective articles after overhaul = $\frac{10}{300} = 0.033$

Since P is not given, we estimate it as

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(400)(0.05) + (300)(0.033)}{700}$$

= 0.0429 and Q = 1 – 0.0429 = 0.9571

$$Z = \frac{0.05 - 0.033}{\sqrt{(0.0429)(0.9571)\left(\frac{1}{400} + \frac{1}{300}\right)}} = 1.0968$$

Table value: The critical value of Z at 1% level of significance is $z_\alpha = 2.33$

Conclusion:

Z= 1.0968 $< Z_\alpha$ = 2.33

Since Calculated value is less than table value. We accept our H_0. Otherwise reject it

∴ We can conclude that there is no significant difference in the improvement of the machine before and after overhaul.

UNIT-V

UNIT-V

SMALL SAMPLE TEST:

When the sample size n is less than 30, i.e., $n < 30$, then the sample is called small sample test.

Exact sample test:

The exact sample tests can, however be applied to large samples also through the converse is not true. In all the exact sample tests, the basic assumption is that "the population from which sample is drawn is normal, i.e., the parent populations are normally distributed".

We consider the following tests under small sample test,

(i) t – test (ii) F – test (iii) $\chi^2 -$ test

Assumptions for student t – test:

- The parent population from which the sample is drawn is normal.
- The sample observations are independent.
- The population S.D of σ is unknown.

t- test for single mean:

Let $x_1, x_2, \ldots x_n$ be a random sample of size n from a normal population with mean μ and variance σ^2. If the sample mean differs significantly from the hypothetical value μ_0 of the population mean.

Null hypothesis (H_0):

There is no significant difference between the sample mean $\bar{x}$ and the population mean μ_0.

Alternative hypothesis (H_1):

There is a significant difference between the sample mean $\bar{x}$ and population mean μ_0.

Test statistic:

Under H_0 , the test statistic is,

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n-1}} \sim t_{(n-1)d.f.}$$

Where $\bar{x} = \frac{\sum x_i}{n}$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Table value:

Find the t table value at the desired level of significance with (n-1)d.f

Conclusion:

Comparing the table value with the calculated value. i.e., ($|t|_{cal} <$ t). If the calculated value > table value we reject H_0 otherwise we accept it.

Problem:1

A random sample of 10 boys had the following I.Q's: 70, 120, 110, 101, 88, 83, 95, 98, 107, and 100. Do these data support the assumption of a population mean I.Q of 100? Find a reasonable in which most of the mean I.Q. values of samples of 10 boys lie.

Given: n = 10, μ = 100

Step 1:

Null hypothesis: (H_0): The data are Consistent with the assumption of a mean I.Q of 100in the population,

i.e., $H_0: \mu = 100$

Step 2:

Alternative hypothesis: (H_1): $H_1: \mu \neq 100$(two – tailed test)

Step 3:

Test statistic: Under H_0, the test statistic is,

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim (t_{n-1})$$

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim (t_{10-1=9})$$

Where $\bar{x} = \frac{\sum x}{n} = \frac{70+120+110+101+88+83+95+98+107+100}{10}$

$$\bar{x} = \frac{972}{10} = 9.72$$

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

$$\bar{x} = 97.2$$

X	70	120	110	101	88	83	95	98	107	100	972
$(x_i - \bar{x})$	-27.2	22.8	12.8	3.8	-9.2	-14.2	-2.2	0.8	9.8	2.8	-
$(x_i - \bar{x})^2$	739.84	519.84	163.84	14.44	84.64	201.64	4.84	0.64	96.04	7.84	1833.6

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

$$= \frac{1}{9} \times 1833.60$$

$s^2 = 203.73$

$s = 14.2734$

$$\therefore\ t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_9 d.f$$

$$= \frac{97.2-100}{14.2734/\sqrt{10}} = \frac{-2.8}{14.2734/3.1623} = \frac{-2.8}{4.5136} = -0.6203$$

$|t| = 0.6203$

Step 4:

Table value: Tabulated $t_{0.05}$ for (10-1), i.e., 9 d.f for two tailed test is 2.262.

Step 5:

Conclusion:

Calculated value = 0.6203

Table value = 2.262

$|t| = 0.6203 < t_\alpha = 2.262$

Since Calculated value is < table value. H_0 is accepted at 5% level of significance and we may conclude that $\mu = 100$.

∴ We can conclude that the data are consistent with the assumption of mean I.Q of 100 in the population.

Problem: 2

The average breaking strength of steel rods is specified to be 18.5 thousand pounds. To test this sample of 14 rods was tested. The mean and s.d obtained were 17.85 and 1.955 respectively. Test the 5% level of significance.

Solution: Given: n = 14; μ = 18.5; $\bar{x}$ = 17.85; S = 1.955

Step 1:

Null hypothesis $H_0: \mu = 18.5$

Step 2:

Alternative hypothesis $H_1: \mu \neq 18.5$

Step 3:

Test statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{n-1}\, d.f$$

$$= \frac{17.85 - 18.5}{1.955/\sqrt{14-1}} \sim t_{14-1}\, d.f = -1.20$$

$|t|$ = 1.20

[∴ Table t at d.f = 13 and α = 5% is equal to 2.16]

Step 4:

Level of significance: 5% or 0.05

Step 5:

Conclusion:

Since, $|t| < 2.16$ so we accept H_0 at 5% level of significance.

Problem: 3

The mean weekly sale of soap bars in departmental stores was 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and should a SD of 17.2 was the advertising campaign successful.

Solution: We are given: n = 22, Population mean $\mu = 146.3$

Sample mean $\bar{x}$ = 153.7, Standard deviation S = 17.2, n – 1 = 22 – 1 = 21

Step 1:

Null hypothesis $H_0: \mu = 146.3$ (or) the advertising campaign is not successful

Step 2:

Alternative hypothesis $H_1: \mu > 146.3$(Right – tailed)

Step 3:

Under H_0 , the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{n-1}\ d.f$$

$$= \frac{153.7 - 146.3}{17.2/\sqrt{22}}$$

$$= \frac{7.4}{17.2/4.6904}$$

$$= \frac{7.4}{3.6671} = 2.0179$$

t = 2.0179

Step 4:

Tabulated $t_{0.05}$ for (22-1), i.e., 21 d .f for right tailed test is 1.721.

Step 5:

Conclusion:

t = 2.0179 > t_α = 1.721

Since calculated value is greater than the table value, we reject our null hypothesis (H_0)**.**
∴ We can conclude that the advertising Campaign is Successful.

Problem: 4

A sample of 26 bulbs gives a mean life of 990 hours with a S.D of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is the sample not the standard?

Solution: Standard given n = 26, $\bar{x} = 990$, $\mu = 1000$, S = 20

Step 1:

Null hypothesis H_0: $\mu = 1000$ the sample is up to the standard.

Step 2:

Alternative hypothesis $H_1: \mu < 1000$ (left – tailed test)

Step 3:

Test statistic $t = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n-1}}} \sim t_{n-1} d.f$

$$|t| = \frac{990 - 1000}{\frac{20}{\sqrt{26-1}}} = +2.5$$

Step 4: Level of significance: 5% or 0.05

Step 5:

Conclusion:

Calculated value = +2.5; Table value = 1.708(left tail test) at 5% level of significance. We reject H_0 if calculated value > table value. The sample is not up to the standard.

Assumptions for student t – test:

The following assumptions are made while applying student t- test:

1. The parent population from while the samples are drawn is normal.
2. The given sample is random. That is, the given sample is drawn by random sampling method.
3. The population standard deviation is not known.

t- test for difference of two means:

Suppose we want to test if two independent samples x_i (i=1, 2, 3, …..n_1) and x_j (j=1, 2, 3, …..n_2) of size n_1 and n_2 have been drawn from two normal populations with mean μ_1 and μ_2 respectively.

Step 1:

Null hypothesis H_0: The samples have been drawn from the normal population with means μ_1 and μ_2.

i.e) $\mu_1 = \mu_2$

Step 2:

Alternative hypothesis H_1: $\mu_1 \neq \mu_2$ ($\mu_1 > \mu_2$ or $\mu_1 < \mu_2$)

Step 3:

The test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} d.f$$

Where $\bar{x}_1 = \frac{\sum x_1}{n_1}$; $\bar{x}_2 = \frac{\sum x_2}{n_2}$; $S^2 = \frac{1}{n_{1+}n_{2-2}} [\sum_i (x_i - \bar{x}_1)^2 + \sum_j (x_i - \bar{x}_2)^2]$

Where $s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$

Step 3:

Level of significance α = 5% or 1%

Find the t table value at the desired l. o. s with $(n_1 + n_2 - 2)$ d.f.

Step 5:

Conclusion:

Since, If $|t_c| < t_\alpha$. H_0 is accepted otherwise H_0 is rejected.

Problem:1

Sample of two types of electric light bulbs were tested for length of life and following data were obtained :

	Type – I	**Type – II**
Sample No:	$n_1 = 8$	$n_2 = 7$
Sample means:	$\bar{x}_1 = 1234\ hrs$	$\bar{x}_2 = 1036\ hrs$
Sample S.D:	$S_1 = 36\ hrs$	$S_2 = 40\ hrs$

Is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life?

Step 1:

Null hypothesis: H_0: $\mu_1 = \mu_2$ i.e. two types I and II of electrical bulbs are identical.

Step 2:

Alternative hypothesis: H_1: $\mu_1 > \mu_2$(Right – tailed test) i.e. type I is superior to type II

Step 3:

Test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} d.f$$

$$S^2 = \frac{1}{n_1+n_2-2}[\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2]$$

$$= \frac{1}{n_1+n_2-2}(n_1 s_1^2 + n_2 s_2^2)$$

$$= \frac{1}{13}[8 \times (36)^2 + 7 \times (40)^2] = 1.65908$$

S = 40.731

$$t = \frac{1234-1036}{40.731\sqrt{\frac{1}{8}+\frac{1}{7}}} = \frac{198}{40.7318\sqrt{0.125+0.143}} = \frac{198}{40.7318\sqrt{0.268}}$$

$$= \frac{198}{21.0869} = 9.3897$$

t = 9.3897

Step 4:

Table value:

Tabulated $t_{0.05}$ for (8+7-2), i.e., 13 d.f. for Right tailed test is 1.771

Step 5:

Conclusion:

t = 9.3897 $>$ t_α = 1.771

Since calculated value is greater than the table value, we reject our null hypothesis (H_0).

∴ We can conclude that type I is superior to type II.

Problem: 2

In a certain experiment to compare two types of animal foods A and B. The following results of increase in weights were observed in animals.

Food A	49	53	51	52	47	50	52	53
Food B	52	55	52	53	50	54	53	

Assuming that the two samples of animals are independent. Can we conclude that food B is better than food A?

Solution:

Step 1:

Null hypothesis $H_0: \mu_1 = \mu_2$. There is no significant difference between population mean and sample mean.

Step 2:

Alternative hypothesis $H_1: \mu_1 < \mu_2$.

Step 3:

Level of significance: $\alpha = 0.05$ or 5%

Step 4:

Test statistic $t = \frac{\bar{x}-\bar{y}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2} d.f$

$$\bar{x} = \frac{\sum x}{n} = 50.875$$

$$\bar{y} = \frac{\sum y}{n} = 52.875$$

$$\sum(x_i - \bar{x})^2 = 30.875 \, , \sum(y_i - \bar{y})^2 = 16.875$$

$$S^2 = \frac{1}{n_1+n_2-2}[\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2]$$

$$= \frac{1}{14}[30.875 + 16.875] = 3.41$$

$$t = \frac{\bar{x}-\bar{y}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2} d.f$$

$$t = \frac{50.875-52.875}{3.41\sqrt{\frac{1}{8}+\frac{1}{8}}}$$

$$t = -2.17$$

Step 5:

Conclusion:

Calculated value = +2.17; Table value = 1.76 at 5% level of significance calculated value > Table value. We reject H_0. then we conclude that food B is better than food A.

Problem: 3

Below are given the gain in weights (in kgs) of pigs fed on two diets A and B

Gain in weight

Diet A	25	32	30	34	24	14	32	24	30	31	35	25	-	-	-
Diet B	44	34	22	10	47	31	40	30	32	35	18	21	35	29	22

Test, if the two diets differ significantly as regards their effect on increase in weight.

Solution:

Step 1: Null hypothesis H_0: there is no significant difference between the mean increase in weight due to diet A and B.

(i.e.,) $H_0: \mu_1 = \mu_2$

Step 2: Alternative hypothesis H_1: $\mu_1 \neq \mu_2$ (two tailed)

Step 3:

Test statistic $$t = \frac{\bar{x}-\bar{y}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2} d.f$$

$$\bar{x} = \frac{\sum x}{n} = \frac{336}{12} = 28$$

$$\bar{y} = \frac{\sum y}{n} = \frac{450}{15} = 30$$

Diet A Diet B

X	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	Y	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
25	-3	9	44	14	196
32	4	16	34	4	16
30	2	4	22	-8	64
34	6	36	10	-20	400
24	-4	16	47	17	289
14	-14	196	31	1	1
32	4	16	40	10	100
24	-4	16	30	0	0
30	2	4	32	2	4
31	3	9	35	5	25
35	7	49	18	-12	144
25	-3	9	21	-9	81
			35	5	25
			29	-1	1
			22	-8	64
336	-	380	450	-	1410

$$\sum(x_i - \bar{x})^2 = 308\,, \qquad \sum(y_i - \bar{y})^2 = 1410$$

$$S^2 = \frac{1}{n_1+n_2-2}[\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2]$$

$$= \frac{1}{12+15-2}[380 + 1410] = 71.6$$

S = 8.4617

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} d.f$$

$$t = \frac{28-30}{8.4617\sqrt{\frac{1}{12}+\frac{1}{15}}} = \frac{-2}{8.4617\sqrt{0.083+0.067}} = \frac{-2}{8.4617\sqrt{0.15}}$$

$$= \frac{-2}{8.4617X0.3873} = -0.6103$$

$|t|$ = 0.6103

Step 4: Test statistic: Tabulated $t_{0.05}$ for (12+15-2) (i.e.,) 25 d.f for two tailed is 2.060

Step 5:

Conclusion:

$|t|$ = 0.6103 $< t_\alpha$ = 2.060

Since Calculated value is less than the table value, we accept our H_0.

∴ We can conclude that there is no significant difference between the mean increase in weight due to Diet A and B.

Assumptions of student t – test for difference of means:

1) Parent population, from which the samples have been drawn are normally distributed.
2) The population variances are equal and unknown, i.e., $\sigma_1{}^2 = \sigma_2{}^2 = \sigma^2(Say)$, where σ^2 is unknown.
3) The two samples are random and independent of each other.

Paired t-test for difference of means:

Let us now consider the case when (i) the sample sizes are equal, (i.e.) $n_1 = n_2$ = n

(say) , and (ii) the two samples are not independent but the sample observations are paired together, (i.e) the pair of observations $(x_{i,}y_i)$, (i=1,2,3,……n) corresponds to the same (i^{th}) sample unit.

Step 1:

Null hypothesis H_0: The sample means differ significantly.

(i.e.,) $H_0 : \mu_1 = \mu_2$

Step 2:

Alternative hypothesis H_1: The sample means are not differ significantly.

(i.e.,) $H_1 : \mu_1 \neq \mu_2$

Step 3:

The test statistic, $t = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1}$ d.f

Where $\bar{d} = \frac{\sum d_i}{n}$; $d_i = x_i - y_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(di - \bar{d})^2$

$$s = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}}$$

Step 4:

Table value: Find the t table value at the desired l.o.s with (n-1) d.f.

Step 5:

Conclusion:

Since, If $|t_c| < t_\alpha$, H_0 is accepted for n – 1 d.f at α% L.O.S otherwise H_0 is rejected.

Problem:1

A certain stimulus administered to each of 12 patients resulted in the following change in blood pressure (bp) 5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, and 6. Can it be concluded that the stimulus will in genera be accompanied by increase in blood pressure.

Solution:

This is a sample of correlated pairs we apply t-test for testing the increase in blood pressure.

Step 1:

Null hypothesis H_0: There is no significant difference in the blood pressure readings of the patients before and after the drug.

(i.e) $H_0 : \mu_1 = \mu_2$

Step 2:

Alternative hypothesis H_1: The stimulus results in an increase in blood pressure .

(i.e) $H_1 : \mu_1 < \mu_2$ (left- tailed test)

Step 3: The test statistic,

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

Patient no	Increase in Bp (d)	d^2
1	5	25
2	2	4
3	8	64
4	-1	1
5	3	9
6	0	0
7	-12	144
8	1	1
9	5	25
10	0	0
11	4	16
12	6	36
Total	$\sum \boldsymbol{d = 31}$	$\sum \boldsymbol{d^2 = 185}$

$$\bar{d} = \frac{\sum d}{n} = \frac{31}{12} = 2.58$$

$$s = \sqrt{\frac{\sum d^2}{12} - \left(\frac{\sum \bar{d}}{n}\right)^2}$$

$$= \sqrt{\frac{185}{12} - (2.58)^2} = 2.96$$

$$t = \frac{\bar{d}}{s/\sqrt{n-1}} \sim t_{n-1} \qquad = \frac{2.58}{2.96/\sqrt{11}} = 2.89 \qquad \text{d . f = n – 1=11}$$

Step 4:

Table value: Tabulated $t_{0.05}$ for (12-1) (i.e.,) 11 d.f for one tailed (left) test is -1.796.

Conclusion:

t = 2.894 $>$ t_α = -1.796. Since calculated value is greater than the table value, we reject our null hypothesis (H_0).

∴ We can conclude that the stimulus will, in general, be accompanied by an increase in blood pressure.

Problem: 2

The weight gains in pounds less than two system of feeding of calves of 10 pairs of identical twins is given below.

Twin pair weight	1	2	3	4	5	6	7	8	9	10
System A	43	39	39	42	46	43	38	44	51	43
System B	37	35	34	41	39	37	37	40	48	56

Discuss whether the difference between two systems of feeding is significant?

Step 1:

Null hypothesis H_0: $\mu_1 = \mu_2$. There is no significant difference between the two systems.

Step 2:

Alternative hypothesis H_1: $\mu_1 \neq \mu_2$. There is a significant difference between the two systems.

Step 3: Level of significance: α = 5% (or) 0.05

Step 4:

Test statistics: Under H_0, the test statistic is,

Twin pair	**System A**	**System B**	$\boldsymbol{d_i}$	$\boldsymbol{(d_i - \bar{d})^2}$
1	43	37	6	2.56
2	39	35	4	0.16
3	39	34	5	0.36
4	42	41	1	11.56
5	46	39	7	6.76
6	43	37	6	2.56
7	38	37	1	11.56
8	44	40	4	0.16
9	51	48	3	1.96
10	43	36	7	6.76
			$\boldsymbol{\sum d_i = 44}$	$\boldsymbol{\sum (d_i - \bar{d})^2 = 44.4}$

$$\bar{d} = \frac{\sum d_i}{n} = \frac{44}{10} = 4.4$$

$$\sum\left(d_i - \bar{d}\right)^2 = 44.4$$

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

$$= \frac{1}{9} \times 44.4 = 4.933$$

S = 2.0817

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{4.4}{2.0817/\sqrt{10}} = 6.684$$

Calculated value = 6.684

Step 5:

Conclusion:

Calculated value = 6.684 > Table value = 2.62 we reject null hypothesis and we conclude that the difference between the two systems is significant.

t- test for correlation coefficient:

If 'r'is the observed correlation in a sample of n pairs of observation from a bivariate normal population.

Step 1:

The null hypothesis H_0: $\rho = 0$ (i.e.,) population correlation coefficient is zero

Step 2:

The alternative hypothesis H_1: $\rho \neq 0$

Step 3:

Level of significance α = 5% or 1% , Find the t table value at the desired l.o.s with (n-1) d.f.

Step 4:

The test statistic is $t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}\ d.f$

Step 5:

Conclusion: Since, If $t_c < t_\alpha$. H_0 is accepted otherwise H_0 is rejected at α% L.O.S

Problem:1

A coefficient of correlation is 0.2 is derived from random samples are 25 pairs of observation. Is this value of r is significant.

Solution:

n = 25, r = 0.2

Step 1:

The null hypothesis H_0: $\rho = 0.2$

Step 2:

The alternative hypothesis H_1: $\rho \neq 0.2$ (two – tailed test)

Step 3:

Level of significance $\alpha = 5\%$

Step 4:

The test statistic is

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

$$t = \frac{0.2\sqrt{25-2}}{\sqrt{1-(0.2)^2}} = 0.97$$

$t_c = 0.97$, $t_\alpha = 1.69$

Step 5:

Conclusion: If $t_c < t_\alpha$. H_0 is accepted otherwise H_0 is rejected at $\alpha = 5\%$ L.O.S

Problem: 2

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?

Solution:

Here n = 27, r = 0.6

Step 1:

The null hypothesis H_0: ρ = 0; i.e.,) the observed sample correlation coefficient is not significant of any correlation in the population.

Step 2:

The alternative hypothesis H_1: ρ ≠ 0 (two – tailed test)

Step 3: The test statistic is

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

$$t = \frac{0.6\sqrt{27-2}}{\sqrt{1-(0.6)^2}}$$

$$= \frac{0.6\ X\ 5}{\sqrt{0.64}} = \frac{3}{0.8} = 3.75$$

Step 4:

Table value: Tabulated $t_{0.05}$ for (27-2) (i.e.,) 25 d.f for two tailed test is 2.060.

Conclusion:

$$t = 3.75 > t_\alpha = 2.060$$

Since calculated value is greater than table value, we reject our H_0.

∴ We can conclude that the observed sample correlation coefficient is significant of any correlation in the population.

F – Test for Equality of Two population variances:

Suppose we want to test (i) whether two independent samples x_i (i= 1,2,…..n_1) and

x_j (j= 1,2,…..n_2) have been drawn from the normal populations with the same variance σ^2

(or) (ii) Whether the two independent estimates of the population variance are homogeneous or not.

Step 1:

The null hypothesis H_0: the population variances are equal

(i.e.,) H_0: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$

Step 2:

The alternative hypothesis H_1:the population variances are not equal.

(i.e.,) H_1: $\sigma_X^2 \neq \sigma_Y^2$

Step 3: The test statistic is,

$$F = \frac{S_X^2}{S_Y^2} \sim F_{(n_1-1, n_2-1)d.f}$$

Where $S_X^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1}(x_i - \overline{x})^2$ and

$S_y^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2}(y_i - \overline{y})^2$. Where n_1 & n_2 are the sizes of the samples drawn from the populations and s_x^2 and s_y^2 are the sample variance.

Step 4:

Table value: Find the F table value at the desired l.o.s with (n_1- 1, n_2- 1) d.f

Conclusion:

Since, If $F_c < F_\alpha$, H_0 is accepted, otherwise H_0 is rejected at α% level of significance.

Problem: 1

Values of a variate in two samples are given below

Sample I	5	6	8	1	12	4	3	9	6	10
Sample II	2	3	6	8	1	10	2	8	-	-

Test the significance of the difference between the two sample means and the two sample variances.

Solution:

Step 1:

The null hypothesis H_0: there is no significant difference between the two population variances are equal

(i.e.,) H_0: $\sigma_X^2 = \sigma_Y^2$

Step 2:

The alternative hypothesis H_1: there is significant differences between the two population variances are not equal.

(i.e.,) H_1: $\sigma_X^2 \neq \sigma_Y^2$(two- tailed)

Step 3:

Test statistic: Under H_0, The test statistic is,

$$F = \frac{S_X^2}{S_Y^2} \sim F_{(n_1-1, n_2-1)d.f}$$

$n_1 = 10$, $n_2 = 8$, $\bar{x} = \frac{64}{10} = 6.4$; , $\bar{y} = \frac{40}{8} = 5$

$S_X^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1}(x_i - \overline{x})^2$; $S_y^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2}(y_i - \overline{y})^2$.

Sample I		**Sample II**	
X	$\mathbf{x^2}$	**Y**	$\mathbf{y^2}$
5	25	2	4
6	36	3	9
8	64	6	36
1	1	8	64
12	144	1	1
4	16	10	100
3	9	2	4
9	81	8	64
6	36	-	-
10	100	-	-
$\sum \boldsymbol{x = 64}$	$\sum \boldsymbol{x^2 = 512}$	$\sum \boldsymbol{y = 40}$	$\sum \boldsymbol{y^2 = 282}$

$$S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1}(x_i - \overline{x})^2 = \sqrt{\frac{\sum x^2}{n_1} - \left(\frac{\sum x}{n_1}\right)^2} = \sqrt{\frac{512}{10} - \left(\frac{64}{12}\right)^2} = 10.24$$

$$S_y^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2}(y_i - \overline{y})^2 = \sqrt{\frac{\sum y^2}{n_2} - \left(\frac{\sum y}{n_2}\right)^2} = \sqrt{\frac{282}{8} - \left(\frac{40}{8}\right)^2} = 10.25$$

$$S_X^2 = \frac{n_1}{n_1 - 1} s_x^2 = \frac{10}{9} \times 10.24 = 11.37$$

$$S_Y^2 = \frac{n_2}{n_2 - 1} s_y^2 = \frac{8}{7} \times 10.25 = 11.75$$

The test statistic is given by

$$F = \frac{S_1^2}{S_2^2} = \frac{11.37}{11.75} = 1.09$$

The table value 7, 9 at 5% 3.29

Conclusion:

Since $F_c < F_\alpha$. H_0 is accepted at 5% level of significance.

Problem: 2

Pumpkins were grown under two experimental conditions. Two random sample of 11 and 9 pumpkins show the sample standard deviations of their weights distributions are normal, test the hypothesis that the true variances are equal, against the alternative they are not, at the 5 % level.

Solution:

Here $n_1 = 11$, $n_2 = 9$, $S_X = 0.8$, $S_Y = 0.5$

Step 1:

The null hypothesis H_0: there is no significant difference between the two population variances are equal

(i.e.,) H_0: $\sigma_X^2 = \sigma_Y^2$

Step 2:

The alternative hypothesis H_1: there is significant differences between the two population variances are not equal.

(i.e.,) H_1: $\sigma_X^2 \neq \sigma_Y^2$(two- tailed)

Step 3: Test statistic: Under H_0, The test statistic is,

$$F = \frac{S_X^2}{S_Y^2} \sim F_{(n_1-1, n_2-1)d.f}$$

$$S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \overline{x})^2$$

$$S_y^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (y_i - \overline{y})^2 .$$

$S_X^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_i - \overline{x})^2$

$(n_1 - 1)\, S_X^2 = \sum_{i=1}^{n_1} (x_i - \overline{x})^2$

$(n_1 - 1)\, S_X^2 = n_1 S_x^2$

$S_X^2 = \left(\frac{n_1}{n_1-1}\right) S_x^2$

$S_y^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (y_i - \overline{y})^2$

$(n_2 - 1)\, S_y^2 = \sum_{i=1}^{n_1} (y_i - \bar{y})^2$

$(n_2 - 1)\, S_y^2 = n_2 S_y^2$

$S_Y^2 = \left(\frac{n_2}{n_2-1}\right) S_Y^2$

$= \left(\frac{11}{11-1}\right)(0.8)^2$ $= \left(\frac{9}{9-1}\right)(0.5)^2$

$= \left(\frac{11}{10}\right)(0.64)$ $= \left(\frac{9}{8}\right)(0.25)$

= 1.1 x 0.64 = 1.125x 0.25

= 0.704 = 0.2813

$$F = \frac{S_X^2}{S_Y^2} = \frac{0.704}{0.2813} = 2.5027$$

Step 4:

Table value:

Tabulated $F_{0.05}\ for$(11- 1, 9- 1) (i.e.,) (10, 8) d.f. for two tailed test is 3.34

Conclusion:

F = 2.5027 $< F_\alpha$ = 3.34

Since Calculated value is less than the table value, we accept ourH_0.

∴ We can conclude that there is no significant difference between two population variances.

Chi-Square test for single variance:

Let $X_1, X_2, ...X_n$ be independent sample taken from normal population with mean μ and variance σ^2.

Null Hypothesis:

H_o: There is no significance difference between the sample variance and population variance (ie) $H_o : \sigma^2 = \sigma_o^2$.

Alternative Hypothesis:

H_1: There is a significance difference between the sample variance and population variance (ie) $H_1: \sigma^2 \neq \sigma_o^2$.

Test statistics:

Under H_o, The test statistic is,

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(x_i-\overline{x})^2}{\sigma_0^2}\right] = \frac{1}{\sigma_0^2}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum x_i)^2)}{n}\right]$$

$$\chi^2 = \frac{ns^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

L. o. s :

Find the χ^2_α table value at the desired l.o.s with (n-1) d.f.

Inference:

If the calculated value is less than the tabulated value. If $\left|\chi^2\right| < \chi^2{}_\alpha$. We accept our H_o. otherwise reject our H_o.

Problem:1

A manufactures of car batteries claims that the life of this batteries are approximately normally distributed S.D of 0.9. if a r.s of 10 batteries have S.D is 1.2 Do you think σ greater than 0.9 use 5%.

Solution:

Given that,

N=10, σ =0.9,s=1.2

Null Hypothesis:

H_o: There is no significance difference between the single variance and population. Variance

(ie) H_o : $\sigma^2 = \sigma_o^2$.

Alternative Hypothesis:

H_1: There is a significance difference between the sample variance and population. Variance

(ie) H_1: $\sigma^2 \neq \sigma_o^2$.

Test Statistic:

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

$=10(1.2)^2/(0.9)^2$

$=17.78$

l.o.s:

$\chi^2{}_\alpha(n-1) = \chi^2{}_{5\%}(n-1)$

n=10,

n-1 = 10-1= 9 d.f

$\chi^2{}_{0.05}(9) = 16.919$

Inference:

If the calculated value is greater than the tabulated value. If $\left|\chi^2\right| < \chi^2{}_\alpha$. Since we conclude that we reject our H_o.

Problem:2

The items of sample at the following value 45, 54, 47, 52, 48, 52, 53, 49, 50. Can this sample we regarded as taken from the population. have been 10 as S.D.

Solution

Given that,

$\sigma = 10$, $n = 9$, $\sigma^2 = 100$

Null Hypothesis:

H_o: there is no significance difference between the sample variance and popl. variance (ie) $H_o : \sigma = \sigma_o^2$.

Alternative Hypothesis:

H_1: ther is a significance difference between the sample variance and popl. variance (ie) $H_1: \sigma \neq \sigma_o^2$.

Test Statistic:

X	X^2
45	2025
54	2916
47	2209
52	2704
48	2304
52	2704
53	2809
49	2401
50	2500
$\sum$ =	$\sum X^2$=
450	22572

$$s = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \sqrt{\frac{22572}{9} - \left(\frac{450}{9}\right)^2} \quad s = 2.828$$

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{9(2.8208)^2}{100} = 0.7197$$

l.o.s:

χ^2_{α} (n-1) = $\chi^2_{0.05}$ (9-1) = $\chi^2_{0.05}$ (8) = 15.507

Inference :

$\left|\chi^2\right| < \chi^2_{\alpha}$. Since we conclude that the calculated value is less than tabulated value we accept our H_o.

Chi-Square test for goodness of fit:

In chi-square test in observed frequencies are denoted by O_i and the expected frequencies are denoted by E_i, then the χ^2 test is

Null Hypothesis:

H_o: the fit is good

Alternative Hypothesis:

H_1: the fit is not good.

Test Statistic:

$$\chi^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] \sim \chi^2_{(n-1)}$$

l.o.s:

if χ^2_{α} is the table value for (n-1) d.f

Inferance:

$\left|\chi^2\right| < \chi^2_{\alpha}$. Since we conclude that the calculated value is less than the tabulated value. We accept our H_o.

Problem:1

The number of auto mobile accident per week in certain community is as follows. 12, 8, 20, 2, 14, 10, 15, 6, 9, 4 are these frequencies in agreement with the belief with that accident condition on the same during this 10 week times.

Solution

Null Hypothesis:

H_o: the fit is good

Alternative Hypothesis:

H_1: the fit is not good.

Test Statistic:

O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
12	10	4	0.4
8	10	4	0.4
20	10	100	10
2	10	64	6.4
14	10	16	1.6
10	10	0	0
15	10	25	2.5
6	10	16	1.6
9	10	1	0.1
4	10	36	3.6
		$\sum ((O_i - E_i)^2 / E_i) = 26.6$	

l.o.s:

χ^2_{α} (n-1) = $\chi^2_{0.05}$ (10-1) = $\chi^2_{0.05}$ (9) = 16.919

Inference:

$\left|\chi^2\right| < \chi^2_{\alpha}$. Since we conclude that the calculated value is greater than the tabulated value. We reject our H_o.

Theorem

Prove that for a 2 x 2 contingency table:

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

Proof:

We know that,2×2 contingency table,

a	b	a+b
c	d	c+d
a+c	b+d	N=a+b+c+d

$$E(a)=\frac{(a+c)(a+b)}{N};\ E(b)=\frac{(a+b)(b+d)}{N};$$

$$E(c)=\frac{(c+d)(a+c)}{N};E(d)=\frac{(c+d)(b+d)}{N}$$

$$\chi^2=\sum_{i=1}^{n}\left[\frac{(O_i-E_i)^2}{E_i}\right]$$

$$\chi^2=\frac{[a-E(a)]^2}{E(a)}+\frac{[b-E(b)]^2}{E(b)}+\frac{[c-E(c)]^2}{E(c)}+\frac{[d-E(d)]^2}{E(d)}$$

$$a-E(a)=a-\frac{(a+b)(a+c)}{N}$$

$$=\frac{aN-(a^2+ac+ab+bc)}{N}$$

$$=\frac{a(a+b+c+d)-a^2-ac-ab-bc}{N}$$

$$=\frac{a^2+ab+ac+ad-a^2-ac-ab-bc}{N}$$

$$[a-E(a)]^2=\frac{(ad-bc)^2}{N^2}$$

Similarly..,

$$[b-E(b)]^2=\frac{(ad-bc)^2}{N^2}$$

$$[c-E(c)]^2=\frac{(ad-bc)^2}{N^2}$$

$$[d-E(d)]^2=\frac{(ad-bc)^2}{N^2}$$

$$\chi^2=\frac{(ad-bc)^2}{N^2}\left\{\frac{1}{E(a)}+\frac{1}{E(b)}+\frac{1}{E(c)}+\frac{1}{E(d)}\right\}$$

$$=\frac{(ad-bc)^2}{N^2}\left\{\frac{1}{\frac{(a+c)(a+b)}{N}}+\frac{1}{\frac{(a+b)(b+d)}{N}}+\frac{1}{\frac{(a+c)(c+d)}{N}}+\frac{1}{\frac{(b+a)(c+d)}{N}}\right\}$$

$$=\frac{(ad-bc)^2}{N^2}\times N\left\{\frac{1}{(a+c)(a+b)}+\frac{1}{(a+b)(b+d)}+\frac{1}{(a+c)(c+d)}+\frac{1}{(b+a)(c+d)}\right\}$$

$$= \frac{(ad-bc)^2}{N}\left\{\frac{(b+d)+(a+c)}{(a+b)(a+c)(b+d)} + \frac{(b+d)+(a+c)}{(a+c)(b+d)(c+d)}\right\}$$

$$= \frac{(ad-bc)^2}{N} \times \left\{\frac{1}{(a+b)(a+c)(b+d)} + \frac{1}{(a+c)(b+d)(c+d)}\right\}$$

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

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

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

χ^2- test

Step 1:

The null hypothesis H_0: There is no significant difference between the observed and the expected frequencies.

Step 2:

The alternative hypothesis H_1: There is significant difference between the observed and the expected frequencies.

Step 3:

Level of significance $\alpha = 5\%$ or 1%

Step 4:

The test statistic is,

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right] \sim \chi^2_{(n-1)} \text{ d.f}$$

Step 5:

Conclusion:

Since, If $\chi^2_c < \chi^2_\alpha$, H_0 is accepted at $\alpha\%$ L.O.S. otherwise H_0 is rejected.

Chi-Square test for Independence of Attributes:

An attributes means a quality or characteristics. Let us consider two attributes A & B. A is divided into two classes. The various cell frequencies can be express in the following table as 2 x 2 contingency table.

	A	B	Total
A	A	B	a+b
B	C	D	c+d
Total	a+c	b+d	N

The expected frequency are given by,

$$E(a)=\frac{(a+c)(a+b)}{N};\ E(b)=\frac{(a+b)(d+b)}{N};\ E(c)=\frac{(c+d)(a+c)}{N};E(d)=\frac{(c+d)(b+d)}{N}$$

Null Hypothesis:

H_o: the attributes are independent

Alternative Hypothesis:

H_1: the attributes are not independent

Test Statistic:

$$\chi^2=\sum\left[\frac{(O_i-E_i)^2}{E_i}\right]\sim\chi^2_{(r-1)(c-1)}$$

Where,

r = No.of rows, c= No.of columns.

Inference :

$\left|\chi^2\right|<\chi^2_\alpha$. Since we conclude that the calculated value is less than the tabulated value.

We accept our H_o. otherwise reject it.

Problem:1

On the basis of information given below. State whether the new treatment is comparatively superior to the conventional frequency.

	Favourable	Not Favourable	Total
New	60	30	90
Conventional	40	70	110

Solution

Null Hypothesis:

H_o: the attributes are independent

Alternative Hypothesis:

H_1: the attributes are not independent

Test Statistic:

$E(60) = \frac{(100 \times 90)}{200} 45$; $E(30) = \frac{(90 \times 100)}{200} = 45$;

$E(40) = \frac{(100 \times 10)}{200} = 55$; $E(70) = \frac{(100 \times 110)}{200} = 55$;

O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
60	45	225	5
30	45	225	5
40	55	225	4.09
70	55	225	4.09
$\sum ((O_i - E_i)^2 / E_i) = 18.18$			

l.o.s:

$\chi^2_{(2-1)(2-1)^{(\alpha)}} = \chi^2_{(1)} = 3.841$

Inference:

If $\left|\chi^2\right| < \chi^2_{\alpha}$, since we conclude that the calculated value is greater than the tabulated value. So, we reject our H_o. therefore the attributes are not independent.

Problem:2

The following data is collected on two characteristics.

	Smokers	**Non - Smokers**
Literate	83	57
Illiterate	45	68

Test whether there is no relation between the habit of smoking and literacy.

Solution:

Step 1:

The null hypothesis H_0: There is no evidence of association between smoking habit and literacy.

Step 2:

The alternative hypothesis H_1: There is evidence of association between smoking habit and literacy.

Step 3: Level of significance $\alpha = 5\%$

Step 4:

The test statistic is,

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right] \sim \chi^2_{(n-1)}$$

Observed frequency table

	Smokers	**Non - Smokers**	**Total**
Literate	83	57	140
Illiterate	45	68	113
Total	128	125	253

Expected frequency table

	Smokers	**Non - Smokers**	**Total**
Literate	71	69	140
Illiterate	57	56	140
Total	128	125	253

$\boldsymbol{O_i}$	$\boldsymbol{E_i}$	$\boldsymbol{(O_i - E_i)^2}$	$\boldsymbol{\frac{(O_i - E_i)^2}{E_i}}$
83	71	144	2.03
57	69	144	2.09
45	57	144	2.53
68	56	144	2.57
			$\sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right] = 9.31$

$$\chi^2 = 9.31$$

$$\chi^2_\alpha = 3.84$$

Step 5:

Conclusion:

Since $\chi^2_c > \chi^2_\alpha$, H_0 is rejected at 5% level of significance.

Application of χ^2 test:

χ^2 Distribution has a large no. of application in statistics.

1. To test the hypothetical value of the popl. Variance $\sigma = \sigma_o^2$.
2. To test the goodness of fit.
3. To test the independent attributes.
4. To test the homogeneity of independent test makes of popl. Variance.
5. To combine various probabilities obtain from independent experiment to given a single test of significance.
6. To test the homogeneity of independent estimates of the popl. Correlation co-efficient.

Properties of t-distribution:

1. The value of t- ranges from minus infinity to plus infinity.
2. The mean of the t- distribution is zero. This is so in case of normal curve also.
3. The variance of t- distribution is greater than one and as the sample size increases it tends to move towards unity.
4. The t-distribution like the standard normal distribution is bell-shaped and symmetrical around mean.
5. The t- distribution is more platy kurtic than the normal distribution.

Application of t-distribution:

The following are some important applications of the t-distribution:

1. Test of hypothesis about the population mean.
2. Test of hypothesis about the difference between two means.
3. Test of hypothesis about the difference between two means with dependent samples.
4. Test of hypothesis about coefficient of correlation.

Applications of F- distribution:

1. F-test for equality of two population variances.
2. F- test for testing the significance of an observed multiple correlation coefficient.
3. F- test for testing the significance of an observed sample correlation ratio.
4. F-test for testing the linearity of regression.
5. F- test for equality of several means.

Characteristics of χ^2-test:

1. Test is based on events or frequencies, where as in theoretical distribution, the test is based on mean and standard deviation.
2. To draw inferences, this test is applied, specially testing the hypothesis but not useful for estimation.
3. The test can be used between the entire set of observed and expected frequencies.
4. For every increase in the number of degree of freedom, a new χ^2distribution is formed.
5. It is a general purpose test and such in highly useful in research.

Assumptions for χ^2 distribution:

1. All the Observations must be independent.
2. All the events must be mutually exclusive.
3. There must be large observations.
4. For comparison purposes, the data must be in original units.
5. The sample data must be drawn at random basis.

Uses of χ^2 distribution:

1. χ^2 - test of goodness of fit.
2. χ^2- as a test of independence
3. χ^2 as a test of homogeneity.

INFERENTIAL STATISTICS

ONE MARK

MULTIPLE CHOICE QUESTIONS

UNIT - I

1. Parameter are those constants which occur in:
   a) Samples b) probability density function c) a formula d) none of the above
2. Estimation of parameters in all scientific investigation is of:
   a) Prime importance b) secondary importance c) no use d) deceptive nature
3. Estimate and estimator are:
   a) Synonyms b) different c) related to population d) none of the above
4. An estimator is considered to be the best if its distribution is:
   a) Continuous b) discrete c) concentrated about the true parameter value
   d) normal.

5. An estimator T_n based on a sample of size n is considered to be the best estimator of θ if:

   a) $P\{|T_n - \theta| < \varepsilon\} \geq P\{|T_n^* - \theta| < \varepsilon\}$ b) $P\{|T_n - \theta| > \varepsilon\} \geq P\{|T_n^* - \theta| > \varepsilon\}$

   c) $P\{|T_n - \theta| < \varepsilon\} = P\{|T_n^* - \theta| < \varepsilon\}$ for all θ d) none of the above

6. An estimator of a parameter function $\tau(\theta)$ is said to be the best if it possesses:

   a) Any two properties of a good estimator b) at least properties of a good estimator

   c) all the properties of a good estimator d) all the above

7. The type of estimates is:

   a) point estimate b) interval estimates c) estimation of confidence region d) all the above

8. If an estimator T_n of population parameter θ converges in probability to θ as n tends to infinity is said to be:

   a) Sufficient b) efficient c) consistent d) unbiased

9. The estimator $\sum X/$ n of population mean are:

   a) an unbiased estimator b) a consistent estimator

   c) both (a) and (b) d) neither (a) nor (b)

10. If $X_1, X_2, X_3, X_4 \ldots\ldots\ldots X_n$ is a random sample from a population N (0, σ^2), the sufficient statistic for σ^2 is:

   a) $\sum X_i$ b) $\sum X_i{}^2$ c) $(\sum X_i)^2$ d) none of the above

11. If $x_1, x_2, x_3 \ldots\ldots\ldots x_n$ be a random sample from a N (μ, σ^2)population, the sufficient statistic for μ is:

   a) $\sum(x_i - \overline{x})$ b) $\bar{x}$/n c) $\sum x_i$ d) $\sum(x_i - \overline{x})2$

12. Factorisation theorem for sufficiency is known as:

a) Rao- Blackwell theorem b) Cramer- Rao theorem

c) Chapman-Robins theorem d) Fisher- Neyman theorem

13. Consistency can specially be named as:

a) Simple consistency b) mean- squared consistency

c) Simple consistency and mean squared consistency both d) all the above

14. Bias of an estimator can be:

a) positive b) negative c) either positive or negative d) always zero

15. If X_1, X_2, X_3, X_4 ……X_n be a random sample from an infinite population where

$S^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2$, the unbiased estimator for the population variance σ^2 is:

a) $\frac{1}{n-1}S^2$ b) $\frac{1}{n}S^2$ c) $\frac{n-1}{n}S^2$ d) $\frac{n}{n-1}S^2$

16. If X_1, X_2, X_3, X_4 ……X_n is a random sample from an infinite population, an estimator for the population variance σ^2 such as:

a) $\frac{1}{n}\sum_i (X_i - \overline{X})^2$ is an unbiased estimator of σ^2 b) $\frac{1}{n}\sum_i (X_i - \overline{X})^2$ is a biased estimator of σ^2
c) $\sum_i (X_i - \overline{X})^2$ is an unbiased estimator of σ^2 d) none of the above

17. Crammer – Rao inequality is based on:

a) Stringent conditions b) mild conditions c) no conditions d) none of the above

18. Regularity conditions of Cramer- Rao inequality are related to:

a) integrability of functions b) differentiability of functions c) both (a) and (b)

d) neither (a) nor (b)

19. Crammer – Rao inequality is valid in case of:

a) continuous variables b) discrete variables c) both (a) and (b) d) neither (a) nor (b)

20. Crammer – Rao inequality was given by them:

a) jointly b) in different years c) in the same year d) none of the above

21. The denominator in the Crammer – Rao inequality is known as:

a) Information limit b) lower bound of the variance

c) upper bound of the variance d) all the above

22. The lower bound for the variance of an estimator T_n under amended regularity conditions of Cramer – Rao was given by:

a) R. A. Fisher b) A. Bhattacharya c) Silverstone d) all the above

23. Another name of best asymptotically normal estimator is:

a) minimum variance unbiased estimator b) best linear unbiased estimator

c) consistent asymptotically normal efficient estimator d) all the above

24. The concepts of consistency, efficiency and sufficiency are due to:

a) J.Neyman b) C.R.Rao c) R.A. Fisher d) J. Berkson

25. The credit of inventing the method of moments for estimating the parameters goes to:

a) R. A. Fisher b) J.Neyman c) laplace d) Karl –Pearson

26. Rao- Blackwell theorem enables us to obtain minimum variance unbiased estimator through:

a) unbiased estimators b) complete statistics c) efficient statistics d) sufficient statistics

27. Minimum Chi-square estimators are:

a) consistent b) asymptotically normal c) efficient d) all the above

28. Minimum Chi-square estimators are not necessarily:

a) efficient b) consistent c) unbiased d) all the above

29. Least square estimators of the parameters of linear model are:

a) unbiased b) BLUE c) UMVU d) all the above

30. A sufficient statistic S= s($x_1, x_2, x_3, \ldots\ldots\ldots x_n$) is said to be complete for a parameter θ if:

a) E_θ (S) = 0 → S= 0 b) E_θ (S) = 1 → S= 1 c)either (a) or (b) d) neither (a) nor (b)

31. Efficiency of sample mean as compared to median as an estimate of the mean of a normal population is:

a) 64 per cent b) 157 per cent c) 317 per cent d) 31.5 per cent

32. If T_n is a consistent estimator of θ , then e^{T_n} is a: a) unbiased estimator of e^θ

b) consistent estimator of e^θ c) MVU estimator of e^θ d) none of the above.

UNIT-II

1. The maximum likelihood estimators are necessarily:
   a) Unbiased b) sufficient c) most efficient d) unique
2. Least square estimators under linear set up are:
   a) Unbiased b) UMVUE's c) BLUE's d) all the above
3. For a random sample from a poisson population P(λ), the maximum likelihood estimate of λ is:
   a) Median b) mode c) geometric mean d) mean
4. For a random sample $(x_1, x_2, x_3, \ldots\ldots x_n)$ from a population N(μ, σ^2), the maximum likelihood estimator of σ^2 is:
   a) $\frac{1}{n}\sum_i (X_i - \overline{X})^2$ b) $\frac{1}{n-1}\sum_i (X_i - \overline{X})^2$ c) $\frac{1}{n}\sum_i (X_i - \mu)^2$ d) $\frac{1}{n-1}\sum_i (X_i - \mu)^2$
5. If the variance of an estimator attains the Crammer- Rao lower bound, the estimator is:
   a) Most efficient b) sufficient c) consistent d) admissible
6. By the method of moments one can estimate:
   a) All constants of a population b) only mean and variance of a distribution
   b) All moments of a population distribution d) all the above
7. If $X_1, X_2, \ldots\ldots X_n$ is a random sample from the population having the density function,

   $$f(x; \theta) = \frac{1}{\sqrt{2\pi\theta}}\ e^{-\frac{1x^2}{2\theta}},$$

   then the maximum likelihood estimators for θ is:

   a) $\sqrt{\sum X_i^2/n}$ b) $\sum X_i^2/n$ c) $\sum X_i^2/\sqrt{n}$
8. If $X_1, X_2, \ldots\ldots X_n$ is a random sample of a population

   $$\frac{1}{\theta\sqrt{2\pi}}\ e^{-\frac{x^2}{2}}\theta^2,$$

   the maximum likelihood estimators for θ is:

   b) $\sqrt{\sum X_i^2/n}$ b) $\sum X_i^2/n$ c) $\sum \frac{X_i^2}{n}$ d) $\sum \frac{X_i}{n}$
9. If T = t$(X_1, X_2, X_3, \ldots\ldots X_n)$ is a sufficient statistics for a parameter θ and the unique MLE $\hat{\theta}$ For θ exists, then

   a) $\hat{\theta}$ = t $(X_1, X_2, X_3, \ldots\ldots X_n)$ b) $\hat{\theta}$ is a function of t

   c) $\hat{\theta}$ is independent of t d) none of the above

UNIT-III

1. The idea of testing of hypothesis was first set forth by:
   a) R. A. Fisher b) J. Neyman c) E. L. Lehman d) A. Wald
2. In 1933, the theory of testing of hypothesis was propounded by:
   a) R. A. Fisher b) J. Neyman c) E. L. Lehman d) Karl Pearson
3. A hypothesis may be classified as:
   a) Simple b) composite c) null d) all the above
4. The hypothesis under test is:
   a) Simple hypothesis b) alternative c) null d) none of the above
5. Whether a test is one-sided or two-sided depends on:
   a) Alternative b) composite c) null d) simple hypothesis
6. A wrong decision about H_0 leads to :
   a) One kind of error b) two kinds of error
   b) three kinds of error d) four kinds of error
7. Power of a test is related to:
   a) Type I error b) Type II error c) types I and II errors both d) none of the above
8. If θ is the true parameter and $\beta(\theta)$ is known as:
   a) Power function b) power of the test
   b) operating characteristic function d) none of the above
9. Level of significance is the probability of:

   a) type I error b) type II error c) nor committing error d) any of the above

10. In terms of type II error β and θ, the true parameter, the function 1- β (θ) is called:

    a) power of the test b) power function c) OC function d) none of the above

11. Out of the two types of error in testing, the more severe error is:

    a) type I error b) type II error

    c) both (a) and (b) are equally severe d) no error is severe.

12. Area of the critical region depends on:

    a) size of type I error b) size of type II error

    c) value of the statistic d) number of observation.

13. Critical region of size α which minimised β amongst all critical regions of size α is called:

    a) power critical region b) minimum critical region

    c) best critical region d) worst critical region.

14. A test based on a test statistic is classified as:

a) randomised test b) non-randomised test c) sequential test d) Bayes test

15. Size of critical region is known as:

a) power of the test b) size of type II error

c) Critical value of the test statistics d) size of the test

16. Degrees of freedom are related to:

a) no. of. Observations in a set b) hypothesis under test

c) no. of. Independent Observations in a set d) none of the above

17. A critical function provides the basis for:

a) accepting H_0 b) rejecting H_0 c) no decision about H_0 d) all the above

18. A test which minimises the power of the test for fixed α is known as:

a) optimum test b) randomised test c) Bayes test d) likelihood ratio test

19. A test T which is a least as powerful as any other test of the same size, is called:

a) best test b) most powerful test c) uniformly most powerful test d) none of the above

20. Neyman- Pearson lemma provides:

a) an unbiased test b) a most powerful test c) an admissible test d) minimax test

21. Type I error is committed when the hypothesis is true but our test ---------- it.

Ans: Rejects

22. Type II errors are made when we accept a null hypothesis which is ----------

Ans: Not true

UNIT-IV

1. Large sample theory is applicable when:
   a) $N > 30$ b) $N < 30$ c) N= 30
2. Standard error of number of success is given by:
   a) $\frac{pq}{n}$ b) $\sqrt{npq}$ c) npq
3. For a two tail test when n is large , the value of Z at 0.05 level of significant is:
   a) 1.645 b) 2.58 c) 1.96
4. For testing $P_1 = P_2$ in a large sample, the proper test is:
   a) t – test b) Z test c) F- test
5. The distribution formed of all possible values of a statistics is called the ----------

Ans: Sampling distribution.

6. Standard error provides an idea about the ---------- of sample.

Ans: Unreliability

7. The standard deviation of sampling distribution is called ----------

Ans: Standard error

8. The mean of sampling distribution of means is equal to the ----------

Ans: Population mean.

9. Standard error of the difference of proportions $(p_1 = p_2)$ in two classes under the hypothesis $H_0: P_1 = P_2$ with usual notations is:

   a) $\sqrt{\hat{p}\hat{q}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$ b) $\sqrt{\hat{p}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$ c) $\hat{p}\hat{q}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ d) $\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$

10. Formula for the Standard error of the difference between proportions $(p_1 = p_2)$ under the e hypothesis $H: P_1 \neq P_2)$ with usual notations is:

    $a)\sqrt{\hat{p}\hat{q}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$ b) $\hat{p}\hat{q}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ c) $\hat{q}\hat{q}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ d) $\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$

11. The formula in general for testing the hypothesis for proportions $H_0: P_1 = P_2$vs. $H_1: P_1 \neq P_2$ Is:

    a) $Z = \frac{p_1-p_2}{s_{(p_1-p_2)}}$ b) $Z = \frac{p_1-p_2}{s^2{}_{(p_1-p_2)}}$ c) $Z = \frac{p_1-p_2}{s_{p_1-s\,p_2)}}$ d) none of the above

UNIT-V

1. Student's t – test is applicable in case of :
   a) Small samples b) for samples of size between 5 and 30
   b) Large samples d) none of the above
2. Student's t-test was invented by:
   a) R. A. Fisher b) G. W. Snedecor c) W. S. Gosset d) W. G. Cochran
3. Student's t-test is applicable only when:
   a) The variate values are independent b) the variate is distributed normally
   b) The sample is not large d) all the above
4. To test $H_0: \mu = \mu_0$ vs. $H_1: \mu > \mu_0$ when the population S.D is known, the appropriate test is:
   a) t-test b) Z-test c) chi-square test d) none of the above
5. To test an hypothesis about proportions of items in a class, the usual test is:
   a) t-test b) F-test c) Z-test d) none of the above
6. The degrees of freedom for statistics –t for paired t-test based on n pairs of observations is:
   a) 2(n-1) b) n-1 c) 2n-1 d) none of the above
7. To test H_0: P = 0.4 vs. H_1: P ≠ 0.4 in binomial population, there are eight persons out of fifteen who favoured a proposal. The value of statistic – Z is:
   a) 5.813 b) 1.08 c) 7.32 d) none of the above
8. The hypothesis that the population variance has a specified value can be tested by:
   a) F-test b) Z-test c) χ^2-test d) None of the above
9. Degrees of freedom for statistics -χ^2in case of contingency table of order (2x2) is
   a) 3 b) 4 c) 2 d) 1
10. The hypothesis $H_0: \sigma_1 = \sigma_2$ vs. $H_1: \sigma_1 > \sigma_2$ can be tested by the statistic:
    a) Z= $\frac{|s_1 - s_2|}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$ b) Z= $\frac{|s_1 - s_2|}{\sqrt{\frac{s_1}{2n_1} + \frac{s_2}{2n_2}}}$ c) $Z = \frac{|s_1 - s_2|}{\sqrt{\frac{{s_1}^2}{2n_1} + \frac{{s_2}^2}{2n_2}}}$ d) none of the above
11. Formula for χ^2 for testing a null hypothesis in a multinomial distribution with usual notations is:
    a) $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ b) $\chi^2 = \sum_{i=1}^{k} \frac{{O_i}^2}{E_i}$ - n c) $\chi^2 = \sum_{i=1}^{k} \frac{{O_i}^2}{np_i}$ – n d) all the above
12. Degrees of freedom for Chi-square in case of contingency table of order (4 x3)are:
    a) 12 b) 9 c) 8 d) 6
13. The degrees of freedom for Chi-square in case of dichotomised frequencies are:
    a) 4 b) 2 c) 1 d) 0
14. An exact test for testing the independent of attributes in a contingency table of order (2 x2) was given by:
    a) Karl Pearson b) Pascal c) Demoivre d) R. A. fisher
15. An exact test for testing the independent of attributes in a contingency table of order (2 x2) is based on the calculation of:
    a) The value of statistics -χ^2 b) probabilities of configurations
    b) The value of statistic – Z d) none of the above
16. Coefficient of contingency is calculated when:
    a) The attributes are independent b) the attributes are associated

C) Both (a) and (b) d) neither (a) and (b)

17. The value of coefficient of contingency lies between:
    a) 0 and ∞ b) 0 and 1 c) 0 to 100 d) -1 and +1
18. The ratio of between sample variance and within sample variance follows:
    a) F-distribution b) χ^2-test c) Z- distribution d) t- distribution
19. Analysis of variance utilises:
    a) F-test b) χ^2-test c) Z-test d) t-test
20. Customarily the large variance in the variance ratio for F- statistic is taken:
    a) In the denominator b) in the numerator c) either way d) none of the above
21. The value of statistic χ^2 is zero if and only if:
    a) $\sum_i O_i = \sum_i E_i$ b) $O_i = E_i$ for all i c) E_i is large d) all the above
22. If all frequencies of classes are same, the value of χ^2 is:
    a) 1 b) ∞ c) zero d) none of the above
23. The range of statistics- χ^2 is:
    a) -1 to +1 b) -∞ to ∞ c) 0 to ∞ d) 0 to 1
24. The range of statistics- t is:
    b) -1 to +1 b) -∞ to ∞ c) 0 to ∞ d) 0 to 1
25. Range of the variance ratio F is:
    c) -1 to +1 b) -∞ to ∞ c) 0 to ∞ d) 0 to 1
26. The best critical region consists of:
    a) Extreme positive values b) extreme negative values
    b) Both (a) and (b) d) neither (a) nor (b)
27. In a contingency table, the expected frequencies are computed under:
    a) Null hypothesis H_0 b) alternative hypothesis H_1
    c) H_0 and H_1 both d) no consideration of hypothesis
28. For testing that a bivariate random sample has come from an uncorrelated population the appropriate test is:
    a) Normal deviate test b) $\chi^2 - test$ c) F-test d) t-test
29. A variable which is used in a contingency table to explain the response variable is known as:
    a) Random variable b) discrete variable c) explanatory variable d) dummy variable
30. Calculated value of chi-square less than its degree of freedom leads of:
    a) Acceptance of H_0 directly b) rejecting of H_0 straightway
    b) No decision about H_0 d) none of the above
31. A zero count in a cell of a contingency table having its expected frequency zero is called:
    a) Structural zero b) random zero c) false zero d) all the above
32. A zero in a cell of a contingency table having a finite expected value is termed as:
    b) Structural zero b) random zero c) false zero d) none of the above

QUESTION BANK

UNIT-I

1. Define: (i) population (ii) sample (iii) parameter
   (iv) unbiased estimator (v) sufficient estimator
2. Write a short notes on interval estimation.
3. State any two properties of estimator.
4. Define: Finite and Infinite population with example.
5. Define: (i) Estimator (ii) statistic (iii) parameter space
6. State and prove invariance properties of consistent estimator.
7. Show that sample variance is a consistent estimator for the population variance of the normal distribution.
8. If T is an unbiased estimator of θ , then show that T^2 is a biased estimator of θ^2, also $\sqrt{T}$ is not an unbiased estimator of $\sqrt{\theta}$.
9. If N(μ, σ^2) show that $\bar{x}$ is an unbiased estimator of population mean μ.
10. State Rao-Blackwell theorem. Also mention its importance.
11. State the regularity conditions for Cramer-Rao inequality.
12. Show that $\frac{\sum x_i(\sum x_i - 1)}{n(n-1)}$ is an unbiased estimator of θ^2 for the sample $x_1, x_2, x_3, \ldots\ldots\ldots x_n$ drawn on x. Which takes the value 1 or 0 with respect to probabilities θ and (1- θ).
13. $x_1, x_2, x_3, \ldots\ldots\ldots x_n$ is a random sample from a normal population N$(\mu, 1)$ show that t= $\frac{1}{n}\sum x_i^2$, is an unbiased estimator of $\mu^2 + 1$.
14. State the properties of consistent estimator.
15. Examine whether the following distribution admits sufficient estimator of the parameter 'θ'. F(x, θ) = $\theta\, x^{\theta-1}$; $0 \le x \le 1$.
16. State and prove Cramer-Rao inequality.
17. State and prove Rao-Blackwell theorem.
18. State and prove sufficient conditions for consistency.
19. If N(μ, σ^2) show that $\bar{x}$ is an unbiased estimator of population mean μ.
20. A random sample $(X_1, X_2, X_3, X_4, X_5)$ of size '5' is drawn from a normal population with unknown mean 'μ'. Consider the following estimators to estimateμ.

    $i)$ $t_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$ ii) $t_1 = \frac{X_1 + X_2}{2} + X_3$

    $iii)$ $t_3 = \frac{2X_1 + X_2 + \lambda X_3}{3}$, where λ is such that t_3 is unbiased estimator of 'μ'

    iv) Find λ, t_1 and t_2 unbiased.

    v) State giving reasons, the estimator which is best among t_1 ,t_2 and t_3.

UNIT-II

1. What do you meant by estimation and MLE.
2. State the conditions for $\hat{\theta}$ to be a MLE.
3. Write any two properties of MLE?
4. Obtain the estimator 'p' in binomial distribution using method of moments.
5. Write down the properties of method of moments?
6. Find the MLE for the parameter λ of a Poisson distribution.
7. Explain Method of moments.
8. Estimate α and β in the case of pearson's Type III distribution by the method of moments.

   $F(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\sqrt{\alpha}} x^{\alpha-1} e^{-\beta x}$, $0 \leq x \leq \infty$.
9. Find the MLE of θ for the density function $f(x, \theta) = \theta\ e^{-\theta x}$.
10. Prove that the MLE of the parameter α of a population having density function: $\frac{2}{\alpha^2}(\alpha - x)$ $0<x<\alpha$, for a sample of unit size is 2x, x being the sample value also that the estimate is biased.
11. In random sampling from Normal population $N(\mu, \sigma^2)$, find the MLE for

(i)μ when σ^2 is known (ii)σ^2 when μ is known (iii) The simultaneous estimation of μ and σ^2.

12. Obtain the MLE'S of α and β for a random sample from the following density $f(x)= y_0\ e^{-\beta(x-\alpha)}$, $\alpha \leq x \leq \infty, \beta > 0$.
13. A random variable X takes the values,0,1,2, with respective probabilities $\frac{\theta}{4N}+\frac{1}{2}\left(1-\frac{\theta}{N}\right), \frac{\theta}{2N}+\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right)$ and $\frac{\theta}{4N}+\frac{1-\alpha}{2}\left(1-\frac{\theta}{N}\right)$, where N is a known number and α, θ are unknown parameters. If 75 independent observations on x yielded the values 0,1,2 with frequencies 27, 38, 10 respectively, estimate θ and α by the method of moments.
14. For the double poisson distribution.

    $P(x) = \frac{1}{2} \frac{e^{-m1}m_1{}^x}{x!} + \frac{1}{2} \frac{e^{-m2}m_2{}^x}{x!}$; x=0,1,2,…..

    Show that the estimators for $m_1\ and\ m_2$ by the method of moments are $\mu_1' \pm \sqrt{\mu_2' - \mu_1' - \mu_1'^2}$.

UNIT-III

1. What is meant by test of hypothesis and statistical hypothesis
2. Define: (i) Simple hypothesis (ii) composite hypothesis
3. Define: (i) Alternative hypothesis and Null hypothesis.
4. Write short notes on Type I and Type II error.
5. Explain Critical Region.
6. Let P be the probability that a coin will fall head in a single toss in order to test H_0 = P = 1/2 and H_1 = P = 3/4. The coin is tossed 5 times and H_0 is rejected if more than 3 heads are obtained. Find the probability of type I error and power of the test.
7. Use the Neyman-Pearson lemma to obtain the best critical region for testing $\theta = \theta_0$ against $\theta = \theta_1 > \theta_0$ in the case of normal population N(θ, σ^2) where σ^2 is known.
8. Show that for the normal distribution with zero mean and variance σ^2, the best critical region for testing $H_0 : \sigma = \sigma_0$ VS $H_1 : \sigma_0 = \sigma_1$ is $\sum x^2 \leq a_\alpha \, for \, \sigma_0 > \sigma_1 \, and \, \sum x^2 \geq b_\alpha for \, \sigma_0 < \sigma_1$.
9. State and prove Neyman-Pearson Lemma.

UNIT-IV

1. Explain sampling distribution and standard error
2. What is meant by test of significance?
3. Define: One –tailed and Two-tailed test.
4. State the formula for testing the difference between sample and population proportions.
5. State large sample test statistic testing population mean.
6. Explain the steps in solving the test of significance.
7. Explain the procedure for testing the equality of two proportions.
8. A sample of 900 members has a mean 3.4 cms and S.D 2.61cms Is this sample came from a large population of mean 3.25cms.
9. Derive the test of significance of difference of two means.
10. In a sample of 1000 people, 540 are rice eaters and 460 are wheat eaters. Can we assume that both rice and wheat eaters are equal. Test at 1% level.
11. Derive the test of significance of single mean.
12. A coin is tossed 10,000 times and it turns up head 5195 times. Discuss whether the coin may be regarded as unbiased one.
13. What are the uses of standard error.
14. In two large populations, there are 30 and 25 percent respectively of blue eyed people. Is this difference likely to be ridden in samples of 1200 and 900 respectively from the two populations.
15. Derive the test for significance of a population.
16. From the following table, test is there any significant difference between the mean.

	Mean	S.D	Size of sample
Sample A	55	10	400
Sample B	57	15	100.

UNIT-V

1. What is meant by test of significance for all small samples?
2. State the formula for testing the significance of observed sample correlation coefficient.
3. Write the test statistic for testing two means using small sample test.
4. State any two uses of Chi-square test.
5. State the conditions for chi-square test for goodness of fit.
6. Give the formula for F- test for equality of two population variances.
7. Explain χ^2 test for goodness of fit.
8. Test the hypothesis that σ= 10 given that S =15 for a random sample of size 50 from a normal population.
9. Explain χ^2 test for independence of attributes.
10. Explain the test procedure for testing equality of two population's variances.
11. Explain t test for testing the significance difference between two populations mean.
12. Explain χ^2 test of testing the significant difference between sample variance and population variance.
13. Two independent samples of 8 and 7 items respectively had the following values. Is the difference between the means of samples significant?

    Sample I: 9 11 13 11 15 9 12 14

    Sample II: 10 12 10 14 9 8 10
14. State and prove 2x2 contingency table.
15. Verify whether the following two samples came from the same population with same variance ($\sigma_1^2 = \sigma_2^2$)

    Sample i: 20 16 26 27 23 22 18 24 25 19

    Sample ii: 27 33 42 35 32 34 38 28 41 43 30 37.