DJS3A - STATISTICAL INFERENCE

Unit - I

Statistical Inference: meaning and purpose, parameter and statistic. Sampling distribution and standard error. Ideal estimator - consistency, unbiasedness, efficiency and sufficient statistic. Unbiased Estimator - Minimum variance unbiased estimator - Cramer-Rao Inequality and Rao-Blackwell theorem.

Unit - II

Point estimation: Moments etimator, maximum liklihood etimator and their properties. Method of Least Squares for regression models. Asymptotic properties of maximum liklihood estimatiors (without proof). Interval estimation for proportions, mean(s), variance(s) based on Chi-square, Student's t, F and normal distributions.

Unit - III

Statistical hypotheses- simple and composite hypotheses-null and alternative hypotheses-critical region- two kinds of errors. Randomized and non-randomized tests -most powerful test-Neyman-Pearson lemma. Likelihood ratio test- tests for mean, equality of two means (independent samples),variance and equality of variances of normal populations.

Unit - IV

Large sample tests concerning mean(s), variance(s), and proportion(s). Exact tests based on t, F and chi-square distributions concerning mean(s), variance(s), correlation coefficient(s). Chi-square Tests: Contingency table, tests for association, independence and goodness of fit.

Unit - V

Non-parametric tests – advantages and disadvantages of nonparametric tests- runs test, Kolmogorov -Smirnov test, sign test, median test, Mann-Whitney U test, and Wilcoxon's signed -rank test.

REFERENCE BOOKS:

1. Rohatgi, V. K. and A. K. md. Ehsanes Saleh (2009) An Introduction to Probability Theory and Mathematical Statistics, 2nd Edition, Wiley Eastern Limited, New Delhi.
2. Gupta, S.C., and V.K. Kapoor (1992) Fundamentals of Mathematical Statistics, A Modern Approach (Eighth Edition). Sultan Chand & sons, New Delhi.
3. Goon, A. M., M.K. Gupta, and B. Dasgupta (2005) Fundamentals of Statistics, Vol. I, (Eigth Edition), World Press, Kolkata.
4. Harold J. Larson (1982) Introduction to Probability Theory and Statistical Inference (Third Edition), John wiley & Sons. Inc., New York.
5. Robert V. Hogg, and Allen T.Craig (1978) Introduction to Mathematical Statistics (Fourth Edition), Macmillan Publishing Co., Inc., New York.
6. Parimal Mukopadhyay (2006) Mathematical Statistics (Third Edition), Books and Allied Private Limited, Kolkata.

STATISTICAL INFERENCE

UNIT-I

1. INTRODUCTION

Statistical inference being similar to, a process of inductive inference as envisaged in classical logic which is the problem is to know the general nature to the phenomenon under study on the basis of the particular set of observations. The only difference is that in a statistical investigation induction is achieved within a probabilistic frame work. Probabilistic considerations enter into the picture in three ways. First, the model used to present the field of study is probabilistic; second, certain probabilistic principles provide the guidelines in making the inference. Third, as we shall see in the sequel, the reliability of the conclusions also judged in probabilistic terms. The problem of statistical inference generally takes one of two forms viz. estimation and hypothesis testing.

POPULATION: The set of all possible observation under study is called population. It is denoted by ‘N’.

PARAMETER: Any population constraints are called parameter. For example, A random variable X~N$\left(\mu,\sigma^2\right)$, here $\mu\ and\ \sigma^2$ are called the parameters of Normal Distribution.

PARAMETRIC SPACE: The set of all possible values of the parameter is called parametric space. (i.e) X~f$\left(x:\theta\right)\forall\ \theta\,\varepsilon\Theta$. It is denoted as ‘$\Theta$’. For example: $X \sim N\left(\mu,\sigma^2\right)\ \forall\ \Theta=\left\{\theta=\left\{\mu,\sigma^2\right\};-\infty<\mu<\infty,\ \sigma>0\right\}$

SAMPLE: It is the subset of the population. It is denoted by ‘n’.

Definition 1: ESTIMATOR

Any function of random samples $x_1,x_2,...,x_n$ that are being observed say $T_n\left(x_1,x_2,...,x_n\right)$ is called an estimator. Clearly a estimator is also a random variable.

If it is used to estimate an unknown parameter, say θ, of the distribution which is also called an estimator.

Definition 2: ESTIMATE

A particular value of the estimator is called estimate of parameter, say θ Eg: $x_1, x_2, ..., x_n$ is a random sample then the mean of the random sample is $\bar{x}$ say T(x)= $\bar{x}$ is called estimator.

1.1 SAMPLING DISTRIBUTION

If a number of samples, each of size n (viz., each sample containing n elements) are drawn from the sample population and if for each sample the values of some statistic say, mean is calculated, a set of values of the statistic will be obtained. These values o the statistic will usually vary from one sample to another, as the values of the population members included in different samples, through drawn from the same population may be different and hence may be treated as values of R.V.

The probability distribution of the statistic (a R.V.) that would be obtained, if the number of samples, each of the size n, were infinitely large, is called the sampling distribution of the statistic. If the random sampling technique is adopted, the nature of the sampling distribution of a statistic can be obtained theoretically, using the theory of probability provided the nature of the population distribution is known.

Like any other distribution, a sampling distribution will have its mean, standard deviation and moments of higher order. The standard deviation of the sampling distribution of a statistic isof particular importance in tests of significance for large samples and testing of hypothesis and is known as standard error (S.E). In the case of large samples (viz. n>30), the sampling distribution of many statistics tend to become normal distributions.

If t is a statistic in large samples, then t follows a normal distribution with mean E(t), which is the corresponding population parameter and S.D. equal to S.E.(t). Hence $Z = \frac{t - E(t)}{S.E.(t)}$ is a standard normal variate.Z follows a normal distribution with mean 0 and S.D. 1 and is called the test statistic.

1.2 STANDARD ERRORS

The standard errors of some frequently occurring statistics for large samples of size n are given below, where σ^2 is the population variance, P, the population proportion and Q=1-P and n$_1$, n$_2$ represent the sizes of two independent random samples drawn from the given population(s).

S.No.	Statistic	Standard Error
1	Sample Mean $(\overline{X})$	$\frac{\sigma}{\sqrt{n}}$
2	Sample Proportion(p)	$\sqrt{PQ/n}$
3	Sample S.D. (s)	$\sqrt{\sigma^2/2n}$
4	Sample Variance (s^2)	$\sigma^2\sqrt{2/n}$
5	Sample Median	$1.25331\,\sigma/\sqrt{n}$
6	Sample coefficient of Variation (υ)	$\frac{\upsilon}{\sqrt{2n}}\sqrt{1+\frac{2\upsilon^3}{10^4}} \approx \frac{\upsilon}{2n}$
7	Sample Correlation Coefficient (r)	$\frac{(1-\rho^2)}{\sqrt{n}}$ where ρ is the population correlation coefficient.
8	Differences of two sample means $(\overline{X}_1 - \overline{X}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
9	Difference of two sample S.D's (s$_1$-s$_2$)	$\sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}$
10	Difference of two sample proportions (p$_1$-p$_2$)	$\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$

Definition 3: ONE PARAMETER EXPONENTIAL FAMILY OF DISTRIBUTION

A random variable $X_1, X_2, \ldots, X_n$ said to be distributed according to a member in one parameter exponential family of distributions if its probability density function is expressed as

$$f(x|\theta) = e^{A(\theta)T(x)-B(\theta)}h(x)\,\forall x \in \aleph, \theta \in \Theta$$

where $A(\theta)$and $B(\theta)$are real valued function of θ, T(x) is a real valued statistic with support x and h(x) is independent of θ.

Definition 4: MULTI PARAMETERS EXPONENTIAL FAMILY OF DISTRIBUTION

A random variables, $X_1, X_2, \ldots, X_n$ which is equal to x_1, x_2,..., x_n is said to be distributed according to a member of multi parameters exponential family of distributions if its probability density function is expressed as

$$f(x_i|\theta) = e^{\sum_{i=1}^{n} A_i(\theta) T_i(x) - B(\theta)} h(x) \forall x \in \aleph, \theta_1, \theta_2 .. \theta_n \in \Theta$$

where $\sum_{i=1}^{n} A_i(\theta)$and $B(\theta)$are real valued function of θ, $\sum_{i=1}^{n} T_i(x)$ is a real valued statistic with support x and h(x) is independent of θ.

Example 1: Let $X_1, X_2, \ldots, X_n$ ~ Poisson(θ). To check whether, this distribution is a member in one parameter exponential family of distribution.

Solution:

$$P(X = x|\theta) = \begin{cases} \dfrac{e^{-\theta}\theta^x}{x!}, x = 0,1,2,\ldots,\theta > 0 \\ 0, \quad otherwise \end{cases}$$

$$P(X = x|\theta) = e^{-\theta}\, e^{x\log\theta} . \frac{1}{x!}$$

$$P(X = x|\theta) = e^{x\log\theta - \theta} . \frac{1}{x!}$$

Here, $A(\theta) = \log\theta; B(\theta) = \theta; T(x) = x; h(x) = \dfrac{1}{x!}$

Therefore, Poisson distribution is a member in one parameter exponential family of distribution.

Example 2: To check whether, the normal distribution is a member in exponential family of distribution.

Solution:

The Probability density function of normal distribution is

$$f\left(x\middle|\mu,\sigma^2\right)=\begin{cases}\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} & ;-\infty<x,\mu<\infty,\sigma>0\\ 0, otherwise\end{cases}$$

$$f\left(x\middle|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2-\log\sigma}$$

$$=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}+\frac{\mu x}{\sigma^2}-\frac{\mu^2}{2\sigma^2}-\log\sigma}$$

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

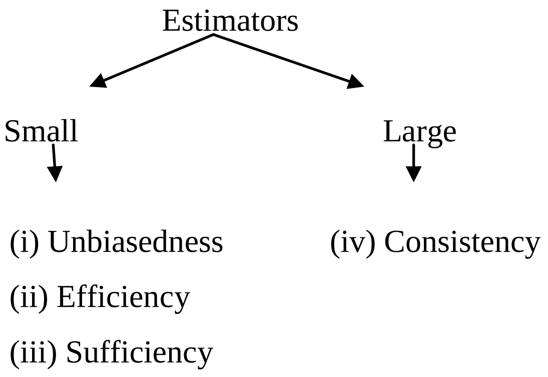
Here, $A(\theta)=\dfrac{-1}{2\sigma^2};B(\theta)=\dfrac{\mu^2}{2\sigma^2}+\log\sigma;T(x_1)=x^2;T(x_1)=x;h(x)=\dfrac{1}{\sqrt{2\pi}}$

Therefore, Normal distribution is a member in multi parameter exponential family of distribution.

1.3 IDEAL/PROPERTIES/ CHARACTERISTICS OF AN ESTIMATOR:

Estimation theory is concerned with the properties of estimators (i.e) with defining properties that can be used to compare different estimator for the same quantity based on the same data. Such properties can be used to determine the best rules to use under given circumstance.

The properties of estimators are mainly classified into two, small sample and large sample properties.



Estimators

Small

(i) Unbiasedness

(ii) Efficiency

(iii) Sufficiency

Large

(iv) Consistency

Definition 5: UNBIASEDNESS

The statistic $T_n = T(x_1, x_2, \ldots, x_n)$ will be called an unbiased estimator of $\gamma(\theta)$ if $E_\theta(T(x)) = \gamma(\theta) \ \ \forall \ \ \theta \ \varepsilon \ \Theta$. (i.e) It has zero bias $\forall \theta$. ($E_\theta(T(x)) - \theta = 0$)

Definition 6: **BIAS**

Let $T_n = T(x_1, x_2, \ldots, x_n)$ is a biased estimator then $E_\theta(T(x)) - \theta = b(\theta)$. Here $b(\theta)$ is amount of bias.

Remarks:

- $E_\theta(T(x)) > \theta$ then bias is positive
- $E_\theta(T(x)) < \theta$ then bias is negative

Mean Square Error: Let $T_n = T(x_1, x_2, \ldots, x_n)$ be an estimator of $\gamma(\theta)$. The mean square error of the estimator $\gamma(\theta)$ is defined as $E_\theta[T - \gamma(\theta)]^2$

(i.e) $E_\theta[T - \gamma(\theta)]^2 = E_\theta[T - \gamma(\theta) + E_\theta(T) - E_\theta(T)]^2$

$$= E_\theta\{[E_\theta(T) - \gamma(\theta)] + [T - E_\theta(T)]\}^2$$

$$= E\left\{(E_\theta(T) - \gamma(\theta))^2 + (T - E_\theta(T))^2 + 2(E_\theta(T) - \gamma(\theta))(T - E_\theta(T))\right\}$$

$$= E\ \{E_\theta(T) - \gamma(\theta)\}^2 + E\{T - E_\theta(T)\}^2$$

$$E_\theta[T - \gamma(\theta)]^2 = b^2\gamma(\theta) + Var(T)$$

$$= \left\{\begin{matrix} bias\, is \\ the\ estimator \end{matrix}\right\} + \left\{\begin{matrix} \text{var}\, iability\ of \\ the\ estimator \end{matrix}\right\}$$

$$=\begin{Bmatrix} accuracy\ of \\ the\ estimator \end{Bmatrix}+\begin{Bmatrix} precision\ of \\ the\ estimator \end{Bmatrix}$$

An estimator is preferred over others if it is a MSE is small as compared to that of others which is achieved by the small variance and small biased both together. Controlling over biased does not necessarily result in low mean square error. Sometimes bearing small amount of biased combined with decrease in variance finally that result into a high decreasing mean square error.

Small mean square error of the estimator results in high probability that the estimator too close to true value of parameter θ by chebyshev's inequality.

The Positive square root of mean square error is called standard error. Mean squared error (MSE) combines the notions of bias and standard error. It is defined as

$$\text{MSE}=(\text{Standard Error})^2+(\text{Bias})^2$$

Example 3: Let $X_1, X_2,...,X_n$ be a random sample from a normal population $\text{N}(\mu,1)$. Show that $t=\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}$ is an unbiased estimator of μ^2+1.

Solution:

Let, $X_1, X_2,...,X_n \sim \text{N}(\mu,1)$, $E(x_i)=\mu$ *and* $V(x_i)=1$

w.k.t; $V(x_i)=E(x_i^{\,2})-[E(x_i)]^2$

$$1 = E(x_i^{\,2})-\mu^2$$

$$E(x_i^{\,2})=1+\mu^2$$

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

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

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

$$=1+\mu^2$$

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

Example 4: If T is an unbiased estimator for θ show that T^2 is an unbiased estimator for θ^2.

Solution:

Since T is an unbiased estimator for θ (i.e) $E(T)=\theta$

w.k.t., $V(T)=E(T^2)-[E(T)]^2$

$$V(T)=E(T^2)-\theta^2$$

$$E(T^2)=V(T)+\theta^2$$

$$E(T^2)\neq\theta^2 \qquad \because V(T)>0\ \left[T^2\ is\,a\,biased\ estimator\ of\ \theta^2\right]$$

Example 5: Show that $\frac{\sum x_i(\sum x_i-1)}{n(n-1)}$ is an unbiased estimate of θ^2 for the sample $x_1, x_2, \ldots, x_n$ drawn an X which takes the value 1 or 0 with respective probabilities $\theta\ and\ (1-\theta)$.

Solution:

$$X \sim Bernoulli(\theta)$$

$$T=\sum x_i \sim Binomial(\theta);\ E(T)=n\theta, V(T)=n\theta(1-\theta)$$

$$E\left[\frac{\sum x_i(\sum x_i-1)}{n(n-1)}\right] = E\left[\frac{T(T-1)}{n(n-1)}\right] = \frac{1}{n(n-1)}E(T^2-T)$$

$$= \frac{1}{n(n-1)}\left[E(T^2)-E(T)\right]$$

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

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

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

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

$$E\left[\frac{\sum x_i(\sum x_i-1)}{n(n-1)}\right] = \theta^2$$

Definition 7: EFFICIENCY

If T_1 is a most efficient estimator with variance V_1 *and* T_2 is any other estimator with variance V_2 then the efficiency of T_2 is defined as $E = \frac{V_1}{V_2}$ obviously E cannot exceed 1. Similarly if T, $T_1, T_2, \ldots, T_n$ are all estimators of $\gamma(\theta)$ and variance of T is minimum then the efficiency E_i *of* $T_i (i = 1,2,\ldots,n)$ is defined as $E_i = \frac{Var\ T}{Var\ T_i}$ for all i=1,2,…,n obviously $E_i \le 1\,(i = 1,2,\ldots,n)$.

Example 6: 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 of estimate μ:

(a) $t_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$

(b) $t_2 = \frac{X_1 + X_2}{2} + X_3$

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

where λ is such that t_3 is an unbiased estimator of μ. Find λ. Are t_1 *and* t_2 unbiased? State giving reasons the estimator which is best among t_1, t_2 *and* t_3.

Solution:

(a) $E\left[\frac{2X_1 + X_2 + \lambda X_3}{3}\right] = \mu$ since t_3 is an unbiased estimator

$$\Rightarrow \frac{1}{3} E[2X_1 + X_2 + \lambda X_3] = \mu$$

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

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

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

$$\Rightarrow \lambda = 0$$

(b) $E(t_1) = E\left[\frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}\right]$

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

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

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

$$= \mu$$

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

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

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

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

$$= 2\mu$$

$\therefore t_1$ *is an unbiased estimator of* μ

t_2 *is a biased estimator of* μ

(c) $V(t_1) = E(t_1{}^2) - [E(t_1)]^2$

$$V(t_1) = V\left[\frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}\right]$$

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

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

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

$$= \frac{1}{4}V[X_1 + X_2] + V[X_3]$$

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

$$= \frac{3\sigma^2}{2}$$

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

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

$$= \frac{5\sigma^2}{9}$$ since $V(t_1) is\ the\ least\ among\ V(t_2), V(t_3)$

$\therefore t_1\ is\ the\ most\ efficient\ estimator\ of\ \mu .$

Example 7: Let X_1, X_2, X_3 is a random sample of size 3 from a population with mean value μ and variance σ^2. T_1, T_2, T_3 are the estimators used to estimate mean value μ, where

$T_1 = X_1 + X_2 - X_3$ $\qquad$ $T_2 = 2X_1 + 3X_3 - 4X_2$ $\qquad$ $T_3 = \frac{1}{3}(\lambda X_1 + X_2 + X_3)$

(i) Are $T_1\ and\ T_2$ unbiased?

(ii) Find the value of $\lambda\ such\ that\ T_3\ is\ ubiased\ estimator\ for\ \mu$

(iii) Which is the most efficient estimator?

(i) $E(T_1) = E(X_1 + X_2 - X_3)$

$= E(X_1) + E(X_2) - E(X_3)$

$= \mu + \mu - \mu$

$= \mu$

$E(T_2) = E(2X_1 + 3X_3 - 4X_2)$

$= 2E(X_1) + 3E(X_3) - 4E(X_2)$

$= 2\mu + 3\mu - 4\mu$

$= \mu$

$\therefore T_1$ is an unbiased estimator of μ

T_2 is an unbiased estimator of μ

(ii) $E(T_3) = \mu$ $\quad$ $\sin ce\ T_3\ is\ an\ unbiased\ estimator\ for\ \mu$

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

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

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

$$\lambda = 1$$

(iii) $V(t_1) = V[X_1 + X_2 - X_3]$

$= \sigma^2 + \sigma^2 + \sigma^2$

$= 3\sigma^2$

$V(t_2) = V[2X_1 + 3X_3 – 4X_2]$

$= 4V(X_1) + 9V(X_3) + 16V(X_2)$

$= 4\sigma^2 + 9\sigma^2 + 16\sigma^2$

$= 29\sigma^2$

$V(t_3) = V\left(\frac{1}{3}(\lambda X_1 + X_2 + X_3)\right)$

$= \frac{1}{9}V[X_1 + X_2 + X_3]$

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

$= \frac{\sigma^2}{3}$

since $V(t_3)$ is the least of all T_1, T_2, T_3

t_3 is the most efficient estimator of μ.

1.4 SUFFICIENCY

Let, the random sample $X_1, X_2, ..., X_n$ have the joint distribution function F_θ which is known expect for k parameters $\theta_1, \theta_2, ..., \theta_k$. We shall write $\theta = (\theta_1, \theta_2, ..., \theta_k)$, a vector with k components, and shall suppose that the parameter space is Θ. Consider k functionally unrelated statistics $T_1, T_2, ..., T_k$, the whole set of which may be denoted that by T.

Definition 8:

Let $X_1, X_2, \ldots, X_n$ be a random sample from the cumulative density function $F_\theta(.) = (F_\theta(.) : \theta \in \Theta)$ where θ is unknown and it is a known family of distribution. A statistic for θ if its conditional distribution of $X_1, X_2, \ldots, X_n$ for any given set of values of $T_1, T_2, \ldots, T_k$ is independent of θ.

Theorem 1: NEYMAN-FISHER FACTORIZATION

Statement:

Let X be a discrete random variable with p.m.f $f(x,\theta)$, $\theta \,\varepsilon \Theta$. Then T(x) is sufficient iff $f(x,\theta) = g(T(x),\theta)h(x) \;\; \forall \, \theta \; \varepsilon \; \Theta$

Proof:

Let $\quad f(x,\theta) = g(T(x),\theta)h(x) \;\; \forall \, \theta \; \varepsilon \; \Theta; \;\; x \varepsilon \Re$

$$= \sum_{x:T(x)=t} g(T(x),\theta)h(x)$$

$$= g(T(x),\theta) \sum_{x:T(x)=t} h(x)$$

Let , $P_\theta(X = x'/T(x) = t) = \begin{cases} 0 & if \;\; T(x') \neq t \\ \dfrac{P_\theta[X = x' \,,\; T(x) = T(x')]}{P[T(x) = T(x')]} & if \;\; T(x') = t \end{cases}$

Consider ,

$$\frac{P_\theta[X = x' \,,\; T(x) = T(x')]}{P[T(x) = T(x')]} = \frac{P_\theta(X = x')}{g(T(x),\theta) \displaystyle\sum_{x:T(x)=t} h(x)}$$

$$= \frac{g(T(x),\theta),\, h(x')}{g(T(x),\theta) \displaystyle\sum_{x:T(x)=t} h(x)}$$

$$= \frac{h(x')}{\displaystyle\sum_{x:T(x)=t} h(x)}$$

$P_\theta(X = x'/T(x)=t)$ = $\dfrac{h(x')}{\sum\limits_{x:T(x)=t} h(x)}$ is independent of θ if T(x')=t.

So the conditional distribution of X given T is independent of the parameter. So T is sufficient statistic for θ.

Conversely, Let T is sufficient for θ.

$$\Rightarrow P_\theta(X = x'/T(x)=t) = \text{C (x', t)} \qquad \because \textit{independent of } \theta$$

$$\Rightarrow \frac{P_\theta[X = x' \text{ , } T(x)=T(x')]}{P[T(x)=T(x')]} = \text{ C (x', t)} \qquad \because T(x')=t$$

$$\Rightarrow P_\theta(X = x') = \text{C (x', t)} \, P[T(x)=T(x')]$$

$$= \text{ C (x', t) } \; g(T(x),\theta)$$

$$\therefore P_\theta(X = x') = g(T(x),\theta)\, h(x)$$

Hence proved.

Example 8:

1. Suppose $X_1, X_2, ..., X_n$ are Independent Identically Distributed (IID) random variables with common probability density function (p.d.f)

$$f(x) = \begin{cases} \theta^x (1-\theta)^{1-x} & if \;\; x = 0,1 \\ 0 & otherwise \end{cases} \quad \text{where } \; 0 < \theta < 1$$

Solution:

The p.d.f of Bernoulli distribution is

$$P(X = x) = \theta^x (1-\theta)^{1-x}$$

The joint probability density function of x_1, x_2,...,x_n is

$$L(\theta) = \prod_{i=1}^{n} P(X_i = x) = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}$$

$$L(\theta)=\left(\frac{\theta}{1-\theta}\right)^{\sum x_i}(1-\theta)^n$$

$$L(\theta)=g\left(\sum x_i,\theta\right)h(.)$$

Here, $g(t)=\left(\frac{\theta}{1-\theta}\right)^{T}(1-\theta)^n;T=\sum_{i=1}^{n}x_i;h(.)=(1-\theta)^n$

Therefore, $\sum x_i$ is a sufficient statistic for θ.

Example 9: Let X ~ Poisson (θ). Find the sufficient statistic for θ.

Solution:

The p.d.f of Poisson distribution is

$$P(X=x)=\begin{cases}\frac{e^{-\theta}\theta^{x}}{x!}, & x=0,1,2\ldots,\theta>0\\ 0, & otherwise\end{cases}$$

$$L(\theta)=\prod_{i=1}^{n}P(X_i=x)=\frac{e^{-n\theta}\theta^{\sum x_i}}{\prod_{i=1}^{n}x_i!}$$

$$L(\theta)=e^{-n\theta}\theta^{\sum x_i}\left(\frac{1}{\prod_{i=1}^{n}x_i!}\right)$$

Here, $g(t)=e^{-n\theta}\theta^{\sum x_i};T=\sum_{i=1}^{n}x_i;h(.)=\left(\frac{1}{\prod_{i=1}^{n}x_i!}\right)$

Therefore, $\sum_{i=1}^{n}x_i$ is a sufficient statistic for θ.

Example 10: Let X ~ Exponential (θ). Find the sufficient statistic for θ.

Solution:

The p.d.f of exponential distribution is

$$f(x|\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$$

$$L(\theta|x) = \frac{1}{\theta^n} e^{-\frac{\sum x_i}{\theta}}$$

$$L(\theta|x) = g\left(\sum_{i=1}^{n} x_i\right) h(.)$$

Here, $g(t) = \frac{1}{\theta^n} e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}}; T = \sum_{i=1}^{n} x_i; h(.) = 1$

Therefore, $\sum_{i=1}^{n} x_i$ is a sufficient statistic for θ.

Example 11: Let X ~ Normal$\left(0, \sigma^2\right)$. Find the sufficient statistic for σ^2.

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

$$f\left(x|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 - \log\sigma}$$

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

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2} - \log\sigma}$$

$$L\left(\sigma^2|x\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\sum x^2}{2\sigma^2} - \log\sigma} = g\left(\sum x_i^2\right) h(.)$$

Here, $g(t)=e^{-\frac{\sum_{i=1}^{n}x_i^2}{2\sigma^2}};T=\sum_{i=1}^{n}x_i^2;h(.)=\left(\frac{1}{\sqrt{2\pi}}\right)^n$

Therefore, $\sum_{i=1}^{n}x_i^2$ is a sufficient statistic for σ^2.

Example 12: Let X ~ Normal $(\mu,1)$. Find the sufficient statistic for μ.

Solution:

$$f(x|\mu,1)=\begin{cases}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu)^2};-\infty<x,\mu<\infty\\ 0, otherwise\end{cases}$$

$$=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(x^2-2x\mu-\mu^2\right)}$$

$$=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}+x\mu-\frac{\mu^2}{2}}$$

$$L(\mu|x)=\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{\mu\sum x_i}\prod_{i=1}^{n}e^{-\frac{1}{2}\left(x_i^2-\mu^2\right)}$$

$$=g\left(\sum x_i,\sum x_i^2\right)\left(\frac{1}{\sqrt{2\pi}}\right)^n$$

Here, $g(t)=e^{\mu\sum x_i}\prod_{i=1}^{n}e^{-\frac{1}{2}\left(x_i^2-\mu^2\right)};T_1=\sum_{i=1}^{n}x_i;T_2=\sum_{i=1}^{n}x_i^2;h(.)=\left(\frac{1}{\sqrt{2\pi}}\right)^n$

Therefore, $\sum_{i=1}^{n}x_i$, $\sum_{i=1}^{n}x_i^2$ is a sufficient statistic for μ.

Example 13: Consider X~f(x: θ), X=1,2,3; $\theta=\theta_1,\theta_2,\theta_3$ with the probability function

x	θ_1	θ_2	θ_3
1	0.1	0.2	0.3
2	0.7	0.4	0.1
3	0.2	0.4	0.6

Show that the statistic $T = \begin{cases} 0 \ \ if \ \ x\, is\, odd \\ 1 \ \ if \ \ x\, is\, even \end{cases}$ is sufficient for θ.

Solution:

The distribution of T is given by

T	θ_1	θ_2	θ_3
0	0.3	0.6	0.9
1	0.7	0.4	0.1

The conditional probability function of x|t is given by

$$P(X|T) = \frac{P(X = x, T = t)}{P(T = t)} = \frac{P(X \cap T)}{P(T)}$$

x	θ_1	θ_2	θ_3
1	1/3	1/3	1/3
2	0	0	0
3	2/3	2/3	2/3

The conditional probability function of x|t when t=1 is given by

x	θ_1	θ_2	θ_3
1	0	0	0
2	1	1	1
3	0	0	0

Since the distribution function of x|t, f(x|t) does not depend on the parameter θ. T is sufficient for θ.

Definition 8: MINIMUM VARIANCE UNBIASED ESTIMATOR

Let $U_{\gamma(\theta)}$ be the class of all unbiased estimator of the parametric function $\gamma(\theta)$. If a statistic $T = T(x_1,..x_n)$ based on sample size n is such that,

(i) T is unbiased for $\gamma(\theta)$ $\forall\ \theta\ \varepsilon\ \Theta$ (i.e) $E(T) = \gamma(\theta)$ $\forall\ \theta\ \varepsilon\ \Theta$

(ii) It has the smallest variance among the class of all unbiased estimators of $\gamma(\theta)$ then T is called minimum variance unbiased estimator of $\gamma(\theta)$. (i.e) $Var(T) \le Var(T^*)$; $T, T^*\ \varepsilon\ U_{\gamma(\theta)}$ *and* $\theta\ \varepsilon\ \Theta$. where T^* is any other unbiased estimator of $\gamma(\theta)$.

Theorem 2:

The minimum variance unbiased estimator is unique in the sense that if T_1 *and* T_2 are minimum variance unbiased estimators for $\gamma(\theta)$ then $T_1 = T_2$ *almost surely*.

Proof:

To prove $T_1 = T_2$

Given T_1 *and* T_2 are unbiased estimator for $\gamma(\theta)$

(i.e) $E(T_1) = E(T_2) = \gamma(\theta) \qquad \forall\ \theta\ \varepsilon\ \Theta$

$Var(T_1) = Var(T_2) \qquad \forall\ \theta\ \varepsilon\ \Theta$

Consider a new estimator $T = \frac{1}{2}(T_1 + T_2)$ which is also unbiased

Since $E(T) = \frac{1}{2}[E(T_1) + E(T_2)]$

$$= \frac{1}{2}[\gamma(\theta) + \gamma(\theta)]$$

$$= \gamma(\theta)$$

$$Var(T) = Var\left[\frac{1}{2}(T_1 + T_2)\right]$$

$$= \frac{1}{4}Var(T_1 + T_2)$$

$$= \frac{1}{4}[Var(T_1) + Var(T_2) + 2Cov(T_1, T_2)]$$

$$= \frac{1}{4}\left[Var(T_1) + Var(T_2) + 2\rho\sqrt{Var(T_1)Var(T_2)}\right]$$

$$= \frac{1}{4}\left[Var(T_1)+Var(T_1)+2\rho\sqrt{Var(T_1)Var(T_1)}\right]$$

$$= \frac{1}{4}\left[2Var(T_1)+2\rho\sqrt{Var^2(T_1)}\right]$$

$$= \frac{1}{4}\left[2Var(T_1)+2\rho Var(T_1)\right]$$

$$= \frac{1}{2}\left[Var(T_1)+\rho Var(T_1)\right]$$

$$= \frac{1}{2}Var(T_1)[1+\rho]$$

where ρ is Karl Pearson's coefficient of correlation between T_1 *and* T_2

Since T_1 is MVUE, $Var(T) \geq Var(T_1)$

$$\Rightarrow \frac{1}{2}Var(T_1)[1+\rho] \geq Var(T_1)$$

$$\Rightarrow \frac{1}{2}[1+\rho] \geq 1$$

$$\Rightarrow \rho \geq 1$$

since $|\rho| \leq 1$ we must have $\rho = 1$

(i.e) T_1 *and* T_2 must have relation of the form:

$$T_1 = \alpha + \beta T_2 \quad \rightarrow 1$$

where α *and* β are constants independent of $x_1, x_2, ..., x_n$ but may depend on θ.

(i.e) we have $\alpha = \alpha(\theta)$ *and* $\beta = \beta(\theta)$

Taking expectation on both sides in equation 1

$$\Rightarrow E(T_1) = E(\alpha) + E(\beta T_2)$$

$$\Rightarrow \theta = \alpha + \beta\theta \quad \rightarrow 2$$

$$Var(T_1) = Var(\alpha + \beta T_2)$$

$$Var(T_1) = \beta^2 Var(T_1)$$

$$\Rightarrow \beta^2 = 1 \qquad \Rightarrow \beta = \pm 1$$

But since $\rho(T_1, T_2) = \pm 1$ the coefficient of regression of T_1 *on* T_2 must be positive, therefore $\beta = 1$

Sub $\beta = 1$ in equation 2

$\Rightarrow \alpha = 0$

Sub α *and* β in equation 1

$\Rightarrow T_1 = T_2$

Hence proved.

Theorem 3:

Let T_1, T_2 be unbiased estimators of $\gamma(\theta)$ with efficiencies ρ_1 *and* ρ_2 respectively and $\rho = \rho_\theta$ be the correlation coefficient between them then

$$\sqrt{\rho_1\rho_2} - \sqrt{(1-\rho_1)(1-\rho_2)} \leq \rho \leq \sqrt{\rho_1\rho_2} + \sqrt{(1-\rho_1)(1-\rho_2)}$$

Proof:

Let T be minimum variance unbiased estimator of $\gamma(\theta)$. Then $E_\theta(T_1) = E_\theta(T_2) = \gamma(\theta)\ \forall\ \theta \varepsilon \Theta$ and

$$\rho_1 = \frac{V_\theta(T)}{V_\theta(T_1)} \qquad \Rightarrow V_\theta(T_1) = \frac{V_\theta(T)}{\rho_1}$$

$$\rho_2 = \frac{V_\theta(T)}{V_\theta(T_2)} \qquad \Rightarrow V_\theta(T_2) = \frac{V_\theta(T)}{\rho_2}$$

Let us consider another estimator $T_3 = \lambda T_1 + \mu T_2$ which is also unbiased estimator of $\gamma(\theta)$.

(i.e) $E(T_3) = E(\lambda T_1 + \mu T_2)$

$= \lambda E(T_1) + \mu E(T_2)$

$= (\lambda + \mu)\gamma(\theta)$

$\Rightarrow \lambda + \mu = 1$

$V_\theta(T_3) = V(\lambda T_1 + \mu T_2)$

$= \lambda^2 V(T_1) + \mu^2 V(T_2) + 2\lambda\mu Cov(T_1, T_2)$

$= \lambda^2 V(T_1) + \mu^2 V(T_2) + 2\rho\lambda\mu\sqrt{Var(T_1)Var(T_2)}$

$$= \lambda^2\frac{V(T)}{\rho_1} + \mu^2\frac{V(T)}{\rho_2} + 2\rho\lambda\mu\sqrt{\frac{V(T)}{\rho_1}\frac{V(T)}{\rho_2}}$$

$$= \lambda^2\frac{V(T)}{\rho_1} + \mu^2\frac{V(T)}{\rho_2} + 2\rho\lambda\mu\, V(T)\frac{1}{\sqrt{\rho_1\rho_2}}$$

$$= \quad V(T)\left[\frac{\lambda^2}{\rho_1}+\frac{\mu^2}{\rho_2}+\frac{2\rho\lambda\mu}{\sqrt{\rho_1\rho_2}}\right]$$

But $V_\theta(T_3) \geq V_\theta(T)$, since $V_\theta(T)$ has minimum variance

$$\Rightarrow V_\theta(T)\left[\frac{\lambda^2}{\rho_1}+\frac{\mu^2}{\rho_2}+\frac{2\rho\lambda\mu}{\sqrt{\rho_1\rho_2}}\right] \geq V_\theta(T)$$

$$\frac{\lambda^2}{\rho_1}+\frac{\mu^2}{\rho_2}+\frac{2\rho\lambda\mu}{\sqrt{\rho_1\rho_2}} \geq 1 = (\lambda+\mu)^2$$

$$\frac{\lambda^2}{\rho_1}+\frac{\mu^2}{\rho_2}+\frac{2\rho\lambda\mu}{\sqrt{\rho_1\rho_2}} \geq \lambda^2+\mu^2+2\lambda\mu$$

$$\left(\frac{\lambda^2}{\rho_1}-\lambda^2\right)+\left(\frac{\mu^2}{\rho_2}-\mu^2\right)+\left(\frac{2\rho\lambda\mu}{\sqrt{\rho_1\rho_2}}-2\lambda\mu\right) \geq 0$$

$$\left(\frac{1}{\rho_1}-1\right)\lambda^2+\left(\frac{1}{\rho_2}-1\right)\mu^2+2\left(\frac{\rho}{\sqrt{\rho_1\rho_2}}-1\right)\mu\lambda \geq 0$$

$$\left(\frac{1}{\rho_1}-1\right)\frac{\lambda^2}{\mu^2}+\left(\frac{1}{\rho_2}-1\right)\frac{\mu^2}{\mu^2}+2\left(\frac{\rho}{\sqrt{\rho_1\rho_2}}-1\right)\frac{\mu\lambda}{\mu^2} \geq 0$$

which is quadratic equation in $\left(\frac{\lambda}{\mu}\right)$

Note that $\rho_i < 1 \Rightarrow \frac{1}{\rho_i} > 1 \quad (or) \ \left(\frac{1}{\rho_i}-1\right) > 0 \ \ \forall\, i = 1,2,\ldots$

We know that ,

$$AX^2+BX+C \geq 0 \ , \quad A>0\ ,\ C>0 \ \text{ iff discriminant is } B^2-4AC \leq 0$$

$$4\left(\frac{\rho}{\sqrt{\rho_1\rho_2}}-1\right)^2-4\left(\frac{1}{\rho_1}-1\right)\left(\frac{1}{\rho_2}-1\right) \leq 0$$

$$\Rightarrow \left(\rho-\sqrt{\rho_1\rho_2}\right)^2-(1-\rho_1)(1-\rho_2) \leq 0$$

$$\Rightarrow \rho^2-2\rho\sqrt{\rho_1\rho_2}+\rho_1\rho_2-1+\rho_2+\rho_1-\rho_1\rho_2 \leq 0$$

$$\Rightarrow \rho^2-2\rho\sqrt{\rho_1\rho_2}+(\rho_1+\rho_2-1) \leq 0$$

$$\rho = \frac{2\sqrt{\rho_1\rho_2} \pm \sqrt{4\rho_1\rho_2-4(\rho_1+\rho_2-1)}}{2}$$

$$= \frac{2\left[\sqrt{\rho_1\rho_2} \pm \sqrt{\rho_1\rho_2 - (\rho_1+\rho_2-1)}\right]}{2}$$

$$= \sqrt{\rho_1\rho_2} \pm \sqrt{(\rho_1-1)(\rho_2-1)}$$

$$\sqrt{\rho_1\rho_2} - \sqrt{(\rho_1-1)(\rho_2-1)} \leq \rho \leq \sqrt{\rho_1\rho_2} + \sqrt{(\rho_1-1)(\rho_2-1)}$$

$$\Rightarrow \sqrt{\rho_1\rho_2} - \sqrt{(1-\rho_1)(1-\rho_2)} \leq \rho \leq \sqrt{\rho_1\rho_2} + \sqrt{(1-\rho_1)(1-\rho_2)}$$

Theorem 4:

If T_1 is a minimum variance unbiased estimator for $\gamma(\theta)$ $\forall\, \theta\, \varepsilon \Theta$ and T_2 is any other unbiased estimator of $\gamma(\theta)$ with efficiency $\rho = \rho_\theta$ then the correlation coefficient between T_1 *and* T_2 is given by $\rho = \sqrt{\rho}$ (i.e) $\rho_\theta = \sqrt{\rho_\theta}$ $\forall\, \theta\, \varepsilon\, \Theta$

Proof:

Using the previous ***Theorem:3*** statement the correlation coefficient ρ lies between

$$\sqrt{\rho_1\rho_2} - \sqrt{(1-\rho_1)(1-\rho_2)} \leq \rho \leq \sqrt{\rho_1\rho_2} + \sqrt{(1-\rho_1)(1-\rho_2)}$$

Here T_1 is a minimum variance unbiased estimator of $\gamma(\theta)$ then the efficiency $\rho_1 = 1$ and T_2 is any other unbiased estimator of $\gamma(\theta)$ with efficiency ρ

$\rho_1 = 1$ *and* $\rho_2 = \rho$ sub in 1

$$\sqrt{1.\rho} \leq \rho \leq \sqrt{\rho}$$

$$\therefore \rho = \sqrt{\rho}$$

Theorem 5:

If T_1 is a minimum variance unbiased estimator for $\gamma(\theta)$ $\forall\, \theta\, \varepsilon \Theta$ and T_2 is any other unbiased estimator of $\gamma(\theta)$ with efficiency $\rho < 1$, then no unbiased linear combination of T_1 *and* T_2 can be an MVUE of $\gamma(\theta)$

Proof:

Consider a linear combination:

$$T = l_1T_1 + l_2T_2$$

will be an unbiased estimator of $\gamma(\theta)$ if

$$E(T) = E(l_1T_1 + l_2T_2) = l_1\, E(T_1) + l_2\, E(T_2) = \gamma(\theta) \quad \forall\, \theta\, \varepsilon\, \Theta$$

$$\Rightarrow l_1 + l_2 = 1 \quad \sin ce \;\; E(T_1) = E(T_2) = \gamma(\theta)$$

The efficiency, $\rho = \dfrac{Var_\theta(T_1)}{Var_\theta(T_2)} \qquad \Rightarrow Var_\theta(T_2) = \dfrac{Var(T_1)}{\rho}$

And $\rho = \rho(T_1, T_2) = \sqrt{\rho}$

$$Var_\theta T = Var_\theta\left[l_1T_1 + l_2T_2\right]$$

$$= l_1^{\,2}\, Var_\theta(T_1) + l_2^{\,2}\, Var_\theta(T_2) + 2l_1l_2\, Cov(T_1, T_2)$$

$$= l_1^{\,2}\, Var_\theta(T_1) + l_2^{\,2}\, Var_\theta(T_2) + 2l_1l_2\, \rho\sqrt{Var(T_1)Var(T_2)}$$

$$= l_1^{\,2}\, Var_\theta(T_1) + l_2^{\,2}\, \frac{Var_\theta(T_1)}{\rho} + 2l_1l_2\, \rho\sqrt{Var(T_1)\frac{Var(T_1)}{\rho}}$$

$$= Var_\theta(T_1)\left[l_1^{\,2} + \frac{l_2^{\,2}}{\rho} + 2l_1l_2\, \frac{\rho}{\sqrt{\rho}}\right]$$

$$= Var_\theta(T_1)\left[l_1^{\,2} + \frac{l_2^{\,2}}{\rho} + 2l_1l_2\right] \qquad \because \rho = \sqrt{\rho}$$

$$Var_\theta(T_1)\left[l_1^{\,2} + l_2^{\,2} + 2l_1l_2\right] < Var_\theta(T) \qquad ; 0 < \rho < 1,\; \frac{1}{\rho} > 1$$

$$Var_\theta(T) > Var_\theta(T_1)(l_1 + l_2)^2$$

$$Var_\theta(T) > Var_\theta(T_1)$$

$\therefore$ T cannot be MVU estimator.

Information Function (Or) Regularity Conditions

(i) Θ is a non degenerate open interval on the head line $\Re$.

(ii) The support of the random variable is independent of the parameter θ.

(iii) $\dfrac{\partial_i\; f(x/\theta)}{\partial\theta_i}$ exists for all i=1,2,3

(iv) $\dfrac{\partial^i}{\partial\theta_i}\displaystyle\int_x f(x/\theta)dx = \int_x \frac{\partial^i\; f(x/\theta)}{\partial\theta_i}dx$ holds for i=1,2,…n

(v) For some function $T(x)$

$$\frac{\partial^i}{\partial\theta_i}\int_x T(x)f(x/\theta)dx = \int_x T(x)\frac{\partial^i\ f(x/\theta)}{\partial\theta_i}dx$$ holds for i=1,2,…n

(vi) $E_\theta\left[\frac{\partial \log\ f_\theta(x_1,x_2,...,x_n)}{\partial_\theta}\right]$ exists and is positive.

It is also called Fisher information measure.

Theorem 6: CRAMER-RAO INEQUALITY

Under the regularity condition if T is an unbiased estimator for $\gamma(\theta)$which is assumed to be a differentiable function of θ satisfies the inequality

$$Var_\theta(T) \geq \frac{[\gamma'(\theta)]^2}{E_\theta\left[\frac{\partial \log\ f_\theta(x_1,x_2,...,x_n)}{\partial\theta}\right]^2}$$ (or)

$$Var_\theta(T) \geq \frac{[\gamma'(\theta)]^2}{I(\theta)}$$

where $I(\theta)$is information measure.

Proof:

Let X be a random variable from the pdf $f(x/\theta)$ and let L be the likelihood function of the random sample $(x_1,x_2,...,x_n)$ from this population. Thus

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

since L is the joint pdf of $(x_1,x_2,...,x_n)$ then

$$\int_x L(\theta/x)dx = 1 \qquad \text{---(1)}$$

where x represents the domain of $(x_1,x_2,...,x_n)$and the integral is an n-fold integral.

Differentiating w.r.t. θ and using regularity conditions, we get

$$\int_x \frac{\partial}{\partial\theta} L\,dx = 0$$

$$\Rightarrow \int_x \left(\frac{\partial}{\partial \theta} \log L \right) L dx = 0 \qquad \text{---(2)}$$

$$\Rightarrow \int_x \left(\frac{1}{L} \right) \frac{\partial L}{\partial \theta} L dx = 0$$

$$\Rightarrow E\left(\frac{\partial}{\partial \theta} \log L \right) = 0 \qquad \forall\ \theta\ \varepsilon\ \Theta$$

Let us consider T(x) = T$\left(x_1, x_2, ..., x_n\right)$ be an unbiased estimator of $\gamma(\theta)$ such that, E(T)= $\gamma(\theta)$.

$$\therefore \int_x T(x) L dx = \gamma(\theta) \qquad \text{---(3)}$$

Differentiating w.r.t. θ we get

$$\int_x T(x) \frac{\partial L}{\partial \theta} dx = \gamma'(\theta)$$

$$\Rightarrow \int_x T(x) \left(\frac{\partial \log L}{\partial \theta} \right) L dx = \gamma'(\theta) \qquad \text{---(4)}$$

Multiplying $\gamma(\theta)$ in Equation 2, we get

$$\int_x \gamma(\theta) \left(\frac{\partial}{\partial \theta} \log L \right) L dx = 0 \qquad \text{---(5)}$$

Subtracting Equation 4 and 5, we get

$$\Rightarrow \int_x [T(x) - \gamma(\theta)] \left(\frac{\partial \log L}{\partial \theta} \right) L dx = \gamma'(\theta) \qquad \text{---(6)}$$

$$\Rightarrow E\left[T(x). \left(\frac{\partial \log L}{\partial \theta} \right) \right] = \gamma'(\theta) \qquad \text{---(7)}$$

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

$$= \gamma'(\theta)$$

We know that,

$$\{Cov(x,y)\}^2 \leq Var(X)Var(Y)$$

$$\Rightarrow [\gamma'(\theta)]^2 \leq Var(T).Var\left(\frac{\partial}{\partial\theta}\log L\right)$$

$$\Rightarrow [\gamma'(\theta)]^2 \leq Var(T)\left\{E\left(\frac{\partial}{\partial\theta}\log L\right)^2 - \left(E\frac{\partial}{\partial\theta}\log L\right)^2\right\}$$

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

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

$$\Rightarrow Var_\theta(T) \geq \frac{[\gamma'(\theta)]^2}{E_\theta\left[\frac{\partial \log f(x_1,x_2,....,x_n)}{\partial\theta}\right]^2} \qquad \text{(or)}$$

$$Var_\theta(T) \geq \frac{[\gamma'(\theta)]^2}{I(\theta)}$$

Hence Proved.

<u>Remarks</u>:

- An unbiased estimator T of $\gamma(\theta)$ for which Cramer-Rao lower bound is attained then it is called minimum variance bound estimator.
- The fisher information measure $I(\theta) = E\left(\frac{\partial}{\partial\theta}\log L\right)^2 = -E\left[\frac{\partial^2}{\partial\theta^2}\log L\right]$

Conditions for the equality sign in Cramer-Rao Inequality

In Cramer-Rao inequality

$$Var_\theta(T) \geq \frac{[\gamma'(\theta)]^2}{E\left(\frac{\partial}{\partial\theta}\log L\right)^2} \qquad \text{---(1)}$$

Rewriting Equation 1, we get

$$Var_\theta(T)E\left(\frac{\partial}{\partial\theta}\log L\right)^2 \geq [\gamma'(\theta)]^2$$

$$\Rightarrow E[T-\gamma(\theta)]^2\ E\left(\frac{\partial}{\partial\theta}\log L\right)^2 \geq [\gamma'(\theta)]^2 \qquad \text{---(2)}$$

The sign of equality will hold in CRR inequality if and only if the sign of equality holds in Equation 2. The sign of equality will hold in Equation 2 by Cauchy-Schwartz inequality $Cov(X,Y)=E(X^2)\ E(Y^2)$ iff the variables $(T-\gamma(\theta))$and $\frac{\partial}{\partial\theta}\log L$ are proportional to each other.

$$\therefore \frac{T-\gamma(\theta)}{\frac{\partial}{\partial\theta}\log L} = \lambda\,(say) = \lambda(\theta)$$

where λ is constant independent of $(x_1,x_2,...,x_n)$ but depend on θ

$$\frac{\partial}{\partial\theta}\log L = \frac{T-\gamma(\theta)}{\lambda(\theta)}$$

$$\Rightarrow T-\gamma(\theta)[A(\theta)] = \frac{\partial}{\partial\theta}\log L \qquad \text{---(3)}$$

where $A=A(\theta)=\frac{1}{\lambda(\theta)}$

Hence the necessary and sufficient condition for an unbiased estimator T to attain the lower bound of the variance is given by Equation 3

Further the C-R minimum variance bound is given by

$$Var\,(T) = \frac{[\gamma'(\theta)]^2}{E\left(\frac{\partial}{\partial\theta}\log L\right)^2} \qquad \text{---(4)}$$

But ,

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

$$= \{A(\theta)\}^2.\,E\{T-\gamma(\theta)\}^2$$

$$= \{A(\theta)\}^2.\,Var(T)$$

Substituting in Equation 4,

$$Var\left(T\right) = \frac{[\gamma'(\theta)]^2}{\{A(\theta)\}^2\, Var(T)}$$

$$\Rightarrow Var(T) = \left|\frac{\gamma'(\theta)}{A(\theta)}\right| = |\gamma'(\theta).\ \lambda(\theta)|$$

Hence , if the likelihood function L is expressible in the form Equation 3 then

1. T is unbiased estimator of $\gamma(\theta)$
2. Minimum variance bound estimator (T) for $\gamma(\theta)$ exists and
3. $Var(T) = \left|\frac{\gamma'(\theta)}{A(\theta)}\right| = |\gamma'(\theta).\ \lambda(\theta)|$

Example 14: Obtain the MVB estimator for μ in normal population $N(\mu, \sigma^2)$ where σ^2 is known.

Solution:

If $x_1, x_2, \ldots, x_n$ is a random sample of size n from the normal population, then

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

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

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

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

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

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

which is of the form

$$\frac{\partial}{\partial\theta}\log L = T - \gamma(\theta)[A(\theta)]$$

then T is a MVB estimator for $\gamma(\theta)$ *and* $A(\theta)$is a constant.

$\therefore \bar{x}$ is a MVB estimator for μ and

$$\Rightarrow Var(\hat{\mu}) = \left|\frac{\gamma'(\theta)}{A(\theta)}\right|$$

$$= \left|\frac{1}{n/\sigma^2}\right|$$

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

Example 15: A random sample $x_1, x_2, ..., x_n$ is taken from the normal population with mean 0 and variance σ^2. Examine if $\sum_{i=1}^{n} x_i^{\,2}/n$ is a MVB estimator for σ^2.

Solution:

If $x_1, x_2, ..., x_n$ is a random sample of size n from the normal population, then

$$L = \prod_{i=1}^{n} f(x_i, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^{n} x_i^{\,2}/2\sigma^2\right\}; \qquad -\infty < x < \infty,\ \sigma > 0$$

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

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

$$\text{Log L} \quad = \quad -n/2\log\sigma^2 - n/2\log 2\pi - \sum_{i=1}^{n} x_i^{\,2}/2\sigma^2$$

$$\frac{\partial}{\partial\sigma^2}\log L = -n/2\sigma^2 - 0 - 1/2\sum_{i=1}^{n} x_i^{\,2}\left(\frac{-1}{\sigma^4}\right)$$

$$= \frac{\sum x_i^{\,2}/n - \sigma^2}{2\sigma^4/n} \text{ which is of the form } T - \gamma(\theta)[A(\theta)]$$

Hence $\hat{\sigma}^2 = \frac{\sum x_i^{\,2}}{n}$ is a MVB estimator and

$$Var(\hat{\sigma}^2) = \left|\frac{\gamma'(\theta)}{A(\theta)}\right|$$

$$= \left|\frac{1}{n/2\sigma^4}\right|$$

$$= \frac{2\sigma^4}{n}$$

Example 16: Find if MVB estimator exists for θ in Cauchy's population

$$\partial F(x;\theta) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2} \quad ; \quad -\infty < x < \infty$$

Solution:

Let $x_1, x_2, ..., x_n$ be a random sample from Cauchy's population.

$$f(x;\theta) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}$$

$$\text{L} = \prod_{i=1}^{n} f(x_i, \theta) = \left(\frac{1}{\pi}\right)^n \prod_{i=1}^{n} \left\{\frac{1}{1+(x-\theta)^2}\right\}$$

$$\log L = -n \log \pi - \sum_{i=1}^{n} \log \left(1+(x_i-\theta)^2\right)$$

$$\frac{\partial}{\partial\theta} \log L = 0 + 2 \sum_{i=1}^{n} \log \frac{(x_i-\theta)}{\left(1+(x_i-\theta)^2\right)}$$

Since $\frac{1}{\pi} \frac{1}{1+(x-\theta)^2}$ cannot be expressed in the form $T-\gamma(\theta)[A(\theta)]$ MVB estimator does not exist for θ in Cauchy's population and so Cramer-Rao lower bound is attainable by the variance of any
unbiased estimator θ.

Example 17: Show that $\overline{X} = \frac{\Sigma x_i}{n}$ in random sampling from

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

is a MVB estimator and has variance $\frac{\sigma^2}{n}$.

Solution:

Let $x_1, x_2, \ldots, x_n$ be a random sample from the population

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

$$\text{L}= \prod_{i=1}^{n} f(x;\theta) = \left(\frac{1}{\theta}\right)^n \ \exp\left[\sum_{i=1}^{n} - x_i / \theta\right]$$

$$\log \text{L} \ = \ -n \log\theta - \sum_{i=1}^{n} x_i / \theta$$

$$\frac{\partial}{\partial\theta} \log L = \frac{-n}{\theta} + \frac{\Sigma x_i}{\theta^2}$$

$$= \ \frac{-n\theta + \Sigma x_i}{\theta^2}$$

$$= \ \frac{n[\Sigma x_i / n \ - \theta]}{\theta^2}$$

$$= \ \frac{\overline{X} - \theta}{\theta^2 / n} \text{ which is of the form } \quad T - \gamma(\theta)[A(\theta)]$$

Hence $\overline{X}$ is the MVB estimator for θ and

$$Var(\hat{\theta}) = \left|\frac{\gamma'(\theta)}{A(\theta)}\right|$$

$$= \left|\frac{1}{1/\theta^2 / n}\right|$$

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

Example 18: Let $x_1, x_2, \ldots, x_n$ be a random sample from a Bernoulli Distribution with parameter p. Then $\theta = p$ and $\Theta = \{\theta \quad 0 < \theta < 1\}$ Find the MVB estimator and its variance.

Solution:

Let $x_1, x_2, \ldots, x_n$ be a random sample from BD.

$$f_\theta(x) = \theta^x (1-\theta)^{1-x}$$

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

$$\log \ L = \Sigma x_i \log\theta + n - \Sigma x_i \log(1-\theta)$$

$$\frac{\partial}{\partial\theta}\log L = \frac{\Sigma x_i}{\theta} + \left[\frac{(n-\Sigma x_i)}{(1-\theta)}\right]$$

$$= \frac{\Sigma x_i}{\theta} - \frac{(n-\Sigma x_i)}{1-\theta}$$

$$= \frac{(1-\theta)\Sigma x_i - \theta(n-\Sigma x_i)}{\theta(1-\theta)}$$

$$= \frac{\Sigma x_i - n\theta}{\theta(1-\theta)}$$

$$= \frac{n[\Sigma x_i / n \ - \ \theta]}{\theta(1-\theta)}$$

$$= \frac{\overline{X} - \theta}{\theta(1-\theta)/n} \text{ which is of the form } \quad T - \gamma(\theta)[A(\theta)]$$

Hence $\overline{X}$ is the MVB estimator for θ and

$$Var(\hat{\theta}) = \left|\frac{\gamma'(\theta)}{A(\theta)}\right|$$

$$= \left|\frac{1}{1/\theta(1-\theta)/n}\right|$$

$$= \frac{\theta(1-\theta)}{n}$$

Example 19: Let $x_1, x_2, ..., x_n$ be a random sample from the Poisson distribution with parameter θ. Find the MVB estimator and its variance.

Solution:

Let $x_1, x_2, ..., x_n$ be a random sample from the Poisson population

$$f_\theta(x) = \frac{e^{-\theta}\,\theta^x}{x!}$$

$$\text{L} = \prod_{i=1}^{n} f(x_i;\theta) = \frac{e^{-n\theta}\ \theta^{\Sigma x_i}}{\prod_{i=1}^{n} x!}$$

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

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

$$= \frac{-n\theta + \Sigma x_i}{\theta}$$

$$= \frac{n[\Sigma x_i / n - \theta]}{\theta}$$

$$= \frac{\overline{X} - \theta}{\theta / n} \text{ which is of the form } T - \gamma(\theta)[A(\theta)]$$

Hence $\overline{X} = \frac{\Sigma x_i}{n}$ is MVB estimator for θ and

$$Var(\hat{\theta}) = \left|\frac{\gamma'(\theta)}{A(\theta)}\right|$$

$$= \left|\frac{1}{1/\theta/n}\right|$$

$$= \frac{\theta}{n}$$

Example 20: Let $x_1, x_2, ..., x_n$ be a random sample from uniform distribution $U(0,\theta)$. Find the MVB estimator and variance.

Solution:

The support of uniform distribution $[U(0,\theta)]$, $\Theta = \{x : 0 < x < \theta\}$ depends on the parameter θ. This violates the regularity conditions and the C-R lower bound theorem does not produce the result.

Completeness:

We discussed one property, viz., sufficiency, that a statistic T may have in relation to a family of distributions. We shall now consider another property, to be called completeness.

Consider the statistic T based on the random variable $X_1, X_2, ..., X_n$ with joint distribution depending on $\theta \in \Theta$. The distribution of T itself will, in general, depend on θ. Hence, related to T, wehave again a family of distributions, say, $\{g(t,\theta),\ \theta\ \varepsilon\ \Theta\}$

Definition 10:

The statistic T=t(x) or more precisely the family of distributions $\{g(t,\theta),\, \theta \,\varepsilon\, \Theta\}$ is said to be complete for θ if

E[h(t)] =0 $\quad \forall\, \theta \quad \Rightarrow P_\theta[h(t)=0]=1$

(i.e) $\int h(t)\, g(t,\theta)dt=0 \quad \forall\, \theta\, \varepsilon \Theta$ (or)

$$\sum_t h(t)\, g(t,\theta)=0 \quad \forall\, \theta\, \varepsilon \Theta$$

$\Rightarrow h(t)=0 \quad \forall\, \theta\, \varepsilon \Theta$ almost surely(a.s)

Definition 11:

The statistic T, or the family of distributions $\{g(t,\theta),\, \theta\, \varepsilon\, \Theta\}$ is said to be boundedly complete for θ if, for any (measurable) function $\psi(T)$ is such that

$|\psi(T)| < M,\ \ for\, some\, M,$

$E_\theta[\psi(T)] = 0\ for\, all\, \theta \in \Theta$

$\Rightarrow \psi(t) = 0\ for\, all\, \theta \in \Theta\ almost\ everywhere$

<u>Note</u>: If T is complete, then it is necessarily boundedly complete.

Theorem 7: RAO-BLACKWELLIZATION

Statement:

Let X and Y are two random variables such that $E(X)=\theta,\ \theta\, \varepsilon\, \Theta$. If a function $\phi(.)$ is defined as $\phi(y)=E(X \mid Y=y)$. Then

(i) $E[\phi(y)] = \theta$ and

(ii) $Var_\theta[\phi(y)] \le Var_\theta(x)$

Proof:

We will give only the proof for the case where the distribution of (x, y) is absolutely continuous.

Let $f_{XY}(x,y)$ denote the joint density function of X and Y.

$f_X(x)$ is the density function of X and

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

To show that $E[\phi(y)] = \theta$

Consider, $\phi(y) = E(X \mid Y = y) = \int_{-\infty}^{\infty} x\, f_{X/Y}(x,y)dx$

Now ,

$$E[\phi(y)] = \int_{-\infty}^{\infty} \phi(y)\, g_Y(y)\, dy$$

$$= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} x\, f_{X/Y}(x,y)dx\right] g_Y(y)\,dy$$

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

where $f_{X/Y}(x,y) = \dfrac{f(x,y)}{g_Y(y)}$

$$E[\phi(y)] = \int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f(x,y)dy\right] dx$$

$$E[\phi(y)] = \int_{-\infty}^{\infty} x\, f_X(x)\; dx \quad = E(x) = \theta$$

Next to show that

$$Var_\theta[\phi(y)] \le Var_\theta(x)$$

Consider, $Var_\theta(x) = E(x-\theta)^2$

$$= E(x - \phi(y) + \phi(y) - \theta)^2$$

$$= E(x-\phi(y))^2 + E(\phi(y)-\theta)^2 + 2E[(x-\phi(y))(\phi(y)-\theta)] \qquad \text{---(1)}$$

Consider,

$$E[(x-\phi(y))(\phi(y)-\theta)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\phi(y))(\phi(y)-\theta)\, f_{X|Y}(x \mid y).\, g_Y(y)dx.dy$$

$$= \int_{-\infty}^{\infty}(\phi(y)-\theta)\left[\int_{-\infty}^{\infty}(x-\phi(y))\, f_{X|Y}(x \mid y)dx\right] g_Y(y)dy$$

$$= \int_{-\infty}^{\infty}(\phi(y)-\theta)\left[\int_{-\infty}^{\infty} x\, f_{X|Y}(x \mid y)dx - \int_{-\infty}^{\infty}\phi(y)\, f_{X|Y}(x \mid y)dx\right] g_Y(y)dy$$

$$= \int_{-\infty}^{\infty} (\phi(y)-\theta)\left[E_\theta(X \mid Y)-\phi(y)\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)dx \right] g_Y(y)dy$$

$$= \int_{-\infty}^{\infty} (\phi(y)-\theta)[E_\theta(X \mid Y)-\phi(y)\ ]\, g_Y(y)dy$$

$$= \int_{-\infty}^{\infty} (\phi(y)-\theta)\,[\phi(y)-\phi(y)\ ]\, g_Y(y)dy$$

$$= 0 \qquad\qquad ---(2)$$

Substitute the Eqn. (2) in Eqn. (1), we get

$$Var_\theta(x) = \ E(x-\phi(y))^2 + V_\theta(\phi(y)) + 0$$

$$\Rightarrow Var_\theta(\phi(y)) = Var_\theta(x) - E(x-\phi(y))^2$$

$$Var_\theta(\phi(y)) \le Var_\theta(x)$$

Hence proved.

Theorem 8: LEHMANN-SCHEFFE

Statement:

If T(X) is a complete sufficient statistic and W(X) is an unbiased estimator of $\tau(\theta)$, then $\phi(T) = E(W/T)$ is an UMVUE of $\gamma(\theta)$. Furthermore $\phi(T)$is the unique UMVUE in the sense that if T* is any other UMVUE, then $P(\phi(T) = T^*) = 1 \quad \forall\ \theta\, \varepsilon\, \Theta$.

Proof:

Let W be any unbiased estimator of $\tau(\theta)$

Then by Rao - blackwell theorem, $\phi(T) = E(W/T)$ is such that $Var_\theta(\phi(T)) \le Var_\theta(W) \quad \forall\ \theta$

Let W* be any other unbiased estimator and

$$\phi^*(T) \ = E[W^*/T] \text{ then}$$

$$E_\theta[\phi(T) - \phi^*(T)] = 0 \quad \forall\ \theta$$

And by the definition of completeness of T, it follows that,

$$P_\theta[\phi(T) = \phi^*(T)] = 1 \ \forall\ \theta$$

Hence , $\phi(T)$is the unique UMVUE.

Definition 12: CONSISTENCY

An estimator $T_n = T(x_1, x_2, ..., x_n)$ based on a random sample of size n is said to be consistent estimator of $\gamma(\theta)$ $\forall\, \theta\, \varepsilon\, \Theta$ if T_n converges to $\gamma(\theta)$in probability (i.e) $T_n \xrightarrow{P} \gamma(\theta)$ as $n \to \infty$. In other words $T_n$is a consistent estimator of $\gamma(\theta)$ if for every $\varepsilon > 0, \eta > 0$there exist a positive integer n which is $\geq m$ such that

$$P[|T_n - \gamma(\theta)| < \varepsilon] \to 1 \quad as\ n \to \infty \quad \forall\, \theta\, \varepsilon\, \Theta$$

$$\Rightarrow P[|T_n - \gamma(\theta)| < \varepsilon] > 1 - \eta \qquad \forall\, (n \geq m)$$

where m is some very large value of n.

Remarks:

If $x_1, x_2, ..., x_n$ is a random sample from population with finite mean $E(x_i) = \mu < \infty$, then by Khinchin’s weak law of large number we have

$$\overline{X}_n = \frac{1}{n} \Sigma x_i \xrightarrow{P} E(X_i) = \mu \quad as\ n \to \infty$$

Hence sample mean $(\overline{X}_n)$ is always a consistent estimator of population mean (μ).

Theorem 9: INVARIANCE PROPERTY OF CONSISTENT ESTIMATOR

Statement:

If $T_n$is a consistent estimator of $\gamma(\theta)$ and $\psi(\gamma(\theta))$is 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 $\gamma(\theta)$

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

Also for every $\varepsilon > 0, \eta > 0$ there exist a positive integer $n \geq m$ such that

$$P\left[\left|T_n - \gamma(\theta)\right| < \varepsilon\right] > 1-\eta \qquad \forall\,(n \geq m) \qquad \text{---(1)}$$

Since $\psi(.)$is a continuous for every $\varepsilon > 0$ however small, there exist a positive number ε_1 such that

$$\left|\psi(T_n) - \psi(\gamma(\theta))\right| < \varepsilon_1 \quad \text{whenever} \quad \left|T_n - \gamma(\theta)\right| < \varepsilon \quad \text{(i.e)} \left|T_n - \gamma(\theta)\right| < \varepsilon$$

$$\Rightarrow \left|\{\psi(T_n) - \psi(\gamma(\theta))\}\right| < \varepsilon_1 \qquad \text{---(2)}$$

For two events A and B if A$\Rightarrow$B then

$$A \subseteq B \quad \Rightarrow P(A) \leq P(B) \qquad \text{(or)}$$

$$P(B) \geq P(B) \qquad \text{---(3)}$$

From Equation 2 and 3, we get

$$P\left[\left|\psi(T_n) - \psi(\gamma(\theta))\right| < \varepsilon_1\right] \geq P\left[\left|T_n - \gamma(\theta)\right| < \varepsilon\right]$$

$$P\left[\left|\psi(T_n) - \psi(\gamma(\theta))\right| < \varepsilon_1\right] \geq 1-\eta \quad \forall\,(n \geq m)$$

$$\Rightarrow \psi(T_n) \xrightarrow{P} \psi(\gamma(\theta)) \quad as \ \ n \to \infty$$

$\therefore \psi(T_n)$ is a consistent estimator of $\psi(\gamma(\theta))$

Theorem 10: SUFFICIENT CONDITION FOR CONSISTENCY

Statement:

Let $\{T_n\}$ be a sequence of estimator such that for all $\theta \ \varepsilon \ \Theta$

1. $E_\theta(T_n) \to \gamma(\theta) \ \ as \ \ n \to \infty$
2. $Var_\theta(T_n) \to 0 \ \ as \ \ n \to \infty$

Then T_n is a consistent estimator of $\gamma(\theta)$.

Proof:

To prove that , T_n is a consistent estimator of $\gamma(\theta)$

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

(i.e) $P[|T_n - \gamma(\theta)| < \varepsilon] > 1 - \eta \qquad \forall\, (n \geq m)$

where ε *and* η are arbitrarily small positive numbers and m is some large value of n.

Applying Chebyshev's inequality to the statistic T_n we get,

$$P[\,|T_n - E_\theta(T_n)| \;\leq \delta\,] \geq 1 - \frac{Var(T_n)}{\delta^2}$$

We have ,

$$|T_n - \gamma(\theta)| \quad = \;[\,|T_n - E_\theta(T_n) + E_\theta(T_n) - \gamma(\theta)|\,]$$

$$\leq |T_n - E_\theta(T_n)| + |E_\theta(T_n) - \gamma(\theta)|$$

Now,

$$|T_n - E_\theta(T_n)| \leq \delta \qquad \Rightarrow |T_n - \gamma(\theta)| \leq \delta + |E(T_n) - \gamma(\theta)|$$

Since for two events A and B if $A \Rightarrow B$ then

$$A \subseteq B \qquad \Rightarrow P(A) \leq P(B) \quad or \quad P(B) \geq P(A)$$

$$P\{|T_n - \gamma(\theta)| \leq \delta + |E_\theta(T_n) - \gamma(\theta)|\} \geq P\{T_n - E\theta(T_n) \leq \delta\}$$

$$\Rightarrow P\{|T_n - \gamma(\theta)| \leq \delta + |E_\theta(T_n) - \gamma(\theta)|\} \geq 1 - \frac{Var_\theta(T_n)}{\delta^2} \qquad \text{---(1)}$$

Given , $E_\theta(T_n) \to \gamma(\theta) \quad \forall\, \theta\, \varepsilon\, \Theta \quad as\; n \to \infty$

Hence for every $\delta_1 > 0$ there exists a particular positive integer $n \geq n_0(\delta_1)$ such that

$$|E_\theta(T_n) - \gamma(\theta)| \leq \delta_1 \qquad n \geq n_0(\delta_1) \qquad \text{---(2)}$$

Also given $Var_\theta(T_n) \to 0 \quad as \;\; n \to \infty$

$$\frac{Var_\theta(T_n)}{\delta_2} \leq \eta \quad \forall \; n \geq n'_0(\eta) \qquad \text{---(3)}$$

where η is arbitrarily small positive number.

Substituting from eqn 2 and 3 in eqn 1 we get,

$$P\left[\left|T_n - \gamma(\theta)\right| \leq \delta + \delta_1\right] \;\; \geq 1 - \eta \quad , \quad n \geq m(\delta_1, \eta)$$

$$P\left[\left|T_n - \gamma(\theta)\right| \leq \varepsilon\right] \;\; \geq 1 - \eta \quad , \; n \geq m$$

where m=max $(n_0 \;, n'_0)$ and $\varepsilon = \delta + \delta_1 > 0$

$$\Rightarrow T_n \xrightarrow{P} \gamma(\theta) \quad as \quad n \to \infty$$

$$\therefore T_n \; is \; a \; consistent \;\; estimator \;\; of \; \gamma(\theta)$$

UNIT-II

In the previous chapter, we have discussed different optimum properties of good point estimators, viz. Unbiasedness, minimum variance, sufficiency, efficiency and consistency. In this chapter, we shall discuss different methods of point estimation which are expected to yield estimators enjoying some of these important properties. Also we shall discuss the confidence interval for proportions, mean(s), variance(s) based on chi-square, Student's t, F and normal distributions.

2.1 METHODS OF ESTIMATION:

There are several methods in estimation theory such as

1. Method of maximum likelihood estimation
2. Method of moments
3. Method of least square
4. Method of minimum variance
5. Method of minimum chi-square
6. Method of inverse probability

METHOD OF MAXIMUM LIKELIHOOD ESTIMATION:

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from population with density function $f(x;\theta)$ then the likelihood function of the sample values $x_1, x_2, \ldots, x_n$ denoted by $L = L(\theta)$ is their joint density function given by

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

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

L gives the relative likelihood that the random variables assume a particular set of values $x_1, x_2, \ldots, x_n$ L becomes a function of a variable θ. The principle of maximum likelihood consist in finding an estimator for the unknown parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ which maximize the likelihood function $L(\theta)$ for variations in parameters (i.e) we want to find $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k\right)$ so that

$$L(\hat{\theta}) > L(\theta) \quad \forall\ \theta\ \varepsilon\, \Theta$$

(i.e) $L(\hat{\theta}) = \sup L(\theta) \quad \forall\ \theta\ \varepsilon\, \Theta$

Thus if there exist a function $\hat{\theta} = \hat{\theta}(x_1, x_2, ..., x_n)$ of the sample values which maximise L for variations in θ. Then $\hat{\theta}$ is to be taken as an estimator of θ. Therefore $\hat{\theta}$ is called maximum likelihood estimator. Thus $\hat{\theta}$ is the solution if any of

$$\frac{\partial L}{\partial \theta} = 0 \quad and \quad \frac{\partial^2 L}{\partial \theta^2} < 0 \qquad \text{---(3)}$$

Since L>0 and logL is a non decreasing function of L. L and logL attain their extreme value (maxima or minima) at the same value of $\hat{\theta}$. Therefore the Equation 2, can be rewritten as

$$\frac{1}{L}\frac{\partial L}{\partial \theta} = 0 \quad \Rightarrow \frac{\partial \log L}{\partial \theta} = 0 \quad and \quad \frac{\partial^2 \log L}{\partial \theta^2} < 0$$

PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATOR:

1. Maximum likelihood estimators are consistent.
2. Any consistent solution of the likelihood equation provides a maximum of the likelihood with probability tends to unity as the sample size tends to unity.
3. Asymptotic normality of MLE: A consistent solution of the likelihood equation is asymptotically normally distributed about the true value of θ_0 (i.e) $\hat{\theta}$ is asymptotically $N\left(\theta_0, \frac{1}{I(\theta_0)}\right)$ $as\ n \to \infty$ where $Var(\hat{\theta}) = \frac{1}{I(\theta)} = \frac{1}{-E\left(\frac{\partial^2}{\partial \theta^2} \log L\right)}$
4. If MLE exist if it is the most efficient in the class of such estimators.
5. If a sufficient estimator exist it is a function of MLE.

6. If for a given population with pdf $f(x:\theta)$ and MVBE T exist for θ then the likelihood equation will have a solution equal to the estimator T.
7. Invariance property of MLE: If T is a MLE of θ and $\psi(\theta)$is a one-to-one function of θ then $\psi(T)$is a MLE of $\psi(\theta)$.

Example 1: In a random sample from normal population $N(\mu,\sigma^2)$ to find the maximum likelihood estimator for the first case (i) μ when $\sigma^2$is known (ii) σ^2 when μ is known.

Solution:

The density function of normal distribution is

$$f(x:\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\sigma^2(x-\mu)^2} \qquad ;-\infty < x,\ \mu < \infty;\sigma > 0$$

Likelihood function is

$$L=\prod_{i=1}^{n} f(x_i : \mu,\sigma^2)$$

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

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

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

Case(i): when $\sigma^2$is known to estimate μ

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

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

$$\frac{\partial^2 \log L}{\partial\mu^2} = \frac{-1}{\sigma^2} < 0$$

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

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

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

$\therefore$ Maximum likelihood estimator for μ when σ^2 is known is a sample mean $\bar{x}$.

Case(ii) : when μ is known to estimate σ^2

$$\frac{\partial \log L}{\partial\sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{\partial^2 \log L}{\partial\sigma^4} = \frac{-n}{2}\left(\frac{-1}{\sigma^4}\right) - \frac{2}{2\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{\partial \log L}{\partial\sigma^2} = 0 \quad \Rightarrow \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2 < 0$$

$$\Rightarrow \frac{-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^4} = 0$$

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

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

$\therefore$ Maximum likelihood function for σ^2 when μ is known is $\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$

Example 2: In a random sample from Poisson distribution with parameter λ. To find maximum likelihood estimator for λ.

Solution:

The Probability density function is

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

The likelihood function is

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

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

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

$$\frac{\partial \log L}{\partial \lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}$$

$$\frac{\partial^2 \log L}{\partial \lambda^2} = \frac{\sum_{i=1}^{n} x_i}{\lambda^2} < 0$$

$$\frac{\partial \log L}{\partial \lambda} = 0 \qquad \Rightarrow\ -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0$$

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

Maximum likelihood estimator for $\lambda\ is\ \frac{\Sigma x_i}{n} = \bar{x}$

Example 3: In a random sample from exponential distribution with parameter θ, find maximum likelihood estimator for θ.

Solution:

The Probability density function is,

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta} \quad ; \qquad 0 < x < \infty,\ \theta > 0$$

The likelihood function is

$$L = \left(\frac{1}{\theta}\right)^n e^{-\sum_{i=1}^{n} x_i / \theta}$$

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

$$\frac{\partial \log L}{\partial \theta} = \frac{-n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i$$

$$\frac{\partial^2 \log L}{\partial \theta^2} = \frac{n}{\theta^2} - \frac{1}{\theta^4} \sum_{i=1}^{n} x_i < 0$$

$$\frac{\partial \log L}{\partial \theta} = \frac{-n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i = 0$$

$$\frac{-n\theta + \Sigma x_i}{\theta^2} = 0$$

$$\hat{\theta} = \frac{\Sigma x_i}{n} = \bar{x}$$

Maximum likelihood estimator for θ is the sample mean $\bar{x}$

2.2 METHOD OF MOMENTS:

This method was discovered by Karl Pearson. Let $f(x : \theta_1, \theta_2, \ldots, \theta_k)$ be the density function of the parent population with k parameters $\theta_1, \theta_2, \ldots, \theta_k$. If μ'_r denotes the rth moment about origin then

$$\mu_r' = \int x^r \; f(x:\theta_1,\theta_2,...,\theta_k)dx \quad \forall\; r = 1,2,...,k \;\rightarrow 1$$

In general $\mu_1',\mu_2',...,\mu_k'$ will be a function of the parameters $\theta_1,\theta_2,...,\theta_k$. Let $x_i \;\; \forall i = 1,2,...,n$ be a random sample of size n from the given population . The method of moments consist in solving the k equations 1 for $\theta_1,\theta_2,...,\theta_k$ in terms of $\mu_1',\mu_2',...,\mu_k'$ and replacing these moments μ'_r for all r = 1,2,..,k by the sample moments.

For example, $\hat{\theta}_i = \theta_i(\hat{\mu}_1',\hat{\mu}_2',...,\hat{\mu}_k')$

$$= \theta_i(m_1',m_2',...,m_k') \qquad \forall\; i = 1,2,...,k$$

where m_i' is the ith moment about the 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.

Example 4: Let X has the following distribution function

X=x	0	1	2
P(X=x)	1-$\theta-\theta^2$	θ	θ^2

Obtain the moment estimate of θ, if in a sample of 25 observations there were 10 one's an 4 two's.

Solutions:

From the given information,

X=x	$P_\theta(X = x)$	Frequency(f)
0	1-$\theta-\theta^2$	11
1	θ	10
2	θ^2	4
Total		25

μ'_1 = E(X) = 0(1-$\theta-\theta^2$) + 1(θ) + 2(θ^2)

$= \theta + 2\theta^2$

$m'_1 = \frac{\Sigma fx}{N} = \frac{18}{25}$

$\mu'_1 = m'_1 \Rightarrow \theta + 2\theta^2 = \frac{18}{25}$

$\Rightarrow 25\theta + 50\theta^2 - 18 = 0$

$\Rightarrow 50\theta^2 + 25\theta - 18 = 0$

$\Rightarrow (10\theta + 9)(5\theta - 2) = 0$

$\Rightarrow \theta = -0.9$ and $\theta = 0.41$

Therefore, the moment estimate of $\theta = 0.41$.

Example 5: A random variable X takes the values 0,1,2 with respective probabilities $\frac{6}{4N} + \frac{1}{2}\left(1 - \frac{\theta}{N}\right)$, $\frac{\theta}{2N} + \frac{\alpha}{2}\left(1 - \frac{\theta}{N}\right)$, $\frac{\theta}{4N} + \frac{1-\alpha}{2}\left(1 - \frac{\theta}{N}\right)$ where N is a known number and α *and* θ are unknown parameters. If 75 independent observations on X give the values 0,1,2 with frequencies 27,38,10 respectively. To estimate, α *and* θ by using method of moments.

Solution:

From the given information,

X=x	$P_\theta(X = x)$	Frequency(f)
0	$\frac{6}{4N} + \frac{1}{2}\left(1 - \frac{\theta}{N}\right)$	27
1	$\frac{\theta}{2N} + \frac{\alpha}{2}\left(1 - \frac{\theta}{N}\right)$	38
2	$\frac{\theta}{4N} + \frac{1-\alpha}{2}\left(1 - \frac{\theta}{N}\right)$	10
Total		75

$$\mu'_1 = \text{E(X)} = 0\left(\frac{6}{4N}+\frac{1}{2}\left(1-\frac{\theta}{N}\right)\right)+1\left(\frac{\theta}{2N}+\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right)\right)+2\left(\frac{\theta}{4N}+\frac{1-\alpha}{2}\left(1-\frac{\theta}{N}\right)\right)$$

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

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

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

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

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

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

$$\mu'_2 = 1\left(\frac{\theta}{2N}+\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right)\right)+4\left(\frac{\theta}{4N}+\frac{1-\alpha}{2}\left(1-\frac{\theta}{N}\right)\right)$$

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

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

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

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

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

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

$$= \quad \frac{3\theta}{2N} + 2 - \frac{3\alpha}{2} - \frac{2\theta}{N} + \frac{3\theta\alpha}{2N}$$

$$= \quad 2 - \frac{\theta}{2N} - \frac{3\alpha}{2}\left[1 - \frac{\theta}{N}\right]$$

$$m'_1 = \frac{\Sigma fx}{N} \quad = \frac{1}{75}\left[0(27) + 1(38) + 2(10)\right]$$

$$= \frac{58}{75}$$

$$m'_2 = \frac{1}{75}\left[0^2(27) + 1^2(38) + 4(10)\right] \qquad = \quad \frac{78}{75}$$

$$\mu'_1 = \; m'_1 \; = \; 1 - \frac{\alpha}{2}\left(1 - \frac{\theta}{N}\right) \; = \frac{58}{75}$$

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

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

$$\mu'_2 \; = \; m'_2 \; = \; 2 - \frac{\theta}{2N} - \frac{3\alpha}{2}\left[1 - \frac{\theta}{N}\right] = \frac{78}{75}$$

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

$$\frac{\theta}{2(75)}+\frac{17}{25} = \frac{150-78}{75}$$

$$\frac{\theta+17(6)}{150} = \frac{72}{75}$$

$$\frac{\theta+102}{150} = \frac{72}{75}$$

$$\theta+102=144$$

$$\hat{\theta}=42$$

Substituting $\theta=42$ in Equation 1,we get

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

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

$$33\alpha=34$$

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

Example 6: To find the moment estimator of Bernoulli population with parameter p.

Solution:

The density function of Bernoulli distribution is

$$P(X=x)=\begin{cases} pq^{1-x} & ; \ x=0\ (or)\ 1,\ 0\le p\le 1\ ,\ p+q=1 \\ 0 & ; \ otherwise \end{cases}$$

Raw moment of Bernoulli distribution

$$\mu'_1 = \text{p}$$

Same moment , $m'_1 = \dfrac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$

The moment estimator is $\mu'_1 = m'_1 \Rightarrow \hat{p} = \bar{x}$

Example 7: To find the moment estimator of Poisson population with parameter λ .

Solution:

$$P(X=x)=p(x)=\begin{cases} \dfrac{e^{-\lambda}\lambda^{x}}{x!} & ; \ x=1,2,\ldots \quad \lambda>0 \\ 0 & ; \ otherwise \end{cases}$$

Since, $\mu'_1 = \lambda$

$$m'_1 = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$\mu'_1 = m'_1 \quad \Rightarrow \bar{x} = \hat{\lambda}$$

Example 8: To find the moment estimator of Exponential distribution with parameter θ .

Solution:

f(x) = $\theta e^{-\theta x}$; $\theta > 0$, $x = 0,1,\ldots$

Since, $\mu'_1 = \theta$

$$m'_1 = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$\mu'_1 = m'_1 \Rightarrow \bar{x} = \hat{\theta}$$

Example 9: To find the moment estimator of Normal distribution with parameter μ *and* σ^2.

Solution:

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

Since, $\mu'_1 = \mu$

$$m'_1 = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$\mu_2' = \mu^2 + \sigma^2$$

$$m'_2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$$

$$\mu'_1 = m'_1 \Rightarrow \hat{\mu} = \bar{x}$$

$$\mu'_2 = m'_2 \Rightarrow \mu^2 + \sigma^2 = \frac{\Sigma \mathrm{x_i}^2}{\mathrm{n}}$$

$$\sigma^2 = \frac{\Sigma \mathrm{x_i}^2}{\mathrm{n}} - \mu^2$$

$$\Rightarrow \hat{\sigma}^2 = \frac{\Sigma x_i^2}{n} - \bar{x}^2$$

2.3 METHOD OF LEAST SQUARES

For fitting a curve of the form

y=f(x; b_0,b_1,...) ---(1)

where b_0,b_1,... are unknown parameters, to the observed sample observations (x_1,y_1), (x_2,y_2),..., (x_n,y_n)by the principle of least squares, we have to minimise

$\sum_i \{y_i - f(x_i; b_0, b_1, b_2, ...)\}^2$ ---(2)

With respect to the parameters $b_0, b_1, ...$.

This is the same as to minimise the sum of squares of the distances of the observed points from the curve measured in the direction of the y-axis.

In case Equation 1 is the regression equation of Y on X, $x_1, x_2, ..., x_n$ may be taken as observed values of the independent variable X, and Y is dependent variable and $e_i = y_i - f(x_i; b_0, b_1, b_2, ...)$ are the residuals or errors. If we assume that the errors are independently normally distributed with zero means and constant variance σ_e^2, then the joint probability density of the errors, or the likelihood function, is given by

$$L = Const.\exp\left[-\frac{1}{2\sigma_e^2}\sum_i \{y_i - f(x_i; b_0, b_1, b_2, ...)\}^2\right]$$

Hence maximising L amounts to minimizing

$$\sum_i \{y_i - f(x_i; b_0, b_1, b_2, ...)\}^2$$

In case e_i's are independently normally distributed with zero means and variances $\sigma_{e_i}^2$, maximizing L will amount to minimizing

$$\sum_i \frac{1}{\sigma_{e_i}^2}\{y_i - f(x_i; b_0, b_1, b_2, ...)\}^2$$

Which is the sum of squares of residuals each weighted by the inverse of its variance. This may be called the weighted least-squares method. In general, we may consider the regression of Y on $X_1, X_2, ..., X_p$ and the method of least squares appropriate for this case may be similarly deduced.

The least-squares estimators do not have any optimum properties even asymptotically. However, in linear estimation this method provides good estimators in small samples.When we are estimating $f(x_i; b_0, b_1, b_2, ...)$ as a linear function of the parameters $b_0, b_1, b_2, ...$, the x_i's being known given values, the least squares estimators obtained as linear functions of the Y's will be minimum-variance unbiased estimators.

Example 10: 1. consider $f(x) = b_0 + b_1 x + b_2 x^2 + ... + b_k x^k, where\, n > k+1.$

Here we have to minimise

$$\sum_i (y_i - b_0 - b_1 x_i - b_2 x_i^2 - ... - b_k x_i^k)^2,$$

with respect to $b_0, b_1, b_2, ..., b_k$. Differentiating this with respect to $b_0, b_1, b_2, ..., b_k$, we have k+1 equations, called the normal equations, given by

$\sum_i x_i^j e_i = 0 \ (j = 0,1,2,...,k)$ or

$$\sum_i x_i^j y_i = b_0 \sum_i x_i^j + b_1 \sum_i x_i^{j+1} + ... + b_k \sum_i x_i^{j+k} \, (j = 0,1,2,...,k)$$

Hence $b_0, b_1, b_2, ..., b_k$ would be obtained as linear functions of the y's.

Example 11: Consider the multiple linear regression $Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p$

Here we have to minimize $\sum \left[y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - ... - b_p x_{pi} \right]^2$ with respect to $b_0, b_1, b_2, ..., b_p$.

The Normal equations are

$$\left. \begin{array}{l} \sum_i e_i = 0 \\ and \ \sum_i x_{ij} e_i = 0 \ (for \ j = 0,1,2,...,p) \end{array} \right\}$$

or

$$\left. \begin{array}{l} \sum_i y_i = nb_0 + b_1 \sum_i x_{1i} + b_2 \sum_i x_{2i} + ... + b_p \sum_i x_{pi} \\ and \ \sum_i x_{ij} y_i = b_0 \sum_i x_{ij} + b_1 \sum_i x_{ji} \, x_{1i} + b_2 \sum_i x_{ji} \, x_{2i} + ... + b_p \sum_i x_{ji} \, x_{pi} \, (j = 0,1,2,...,k) \end{array} \right\}$$

and hence $b_0, b_1, b_2, ..., b_p$ may be obtained as linear functions of y$_i$'s and of the given known values x's.

Definition 1: CONFIDENCE INTERVAL AND LIMITS

Let $x_1, x_2, ..., x_n$ be a random sample from the density $f(., \theta)$. Let $T_1 = t_1(x_1, x_2, ..., x_n)$ and $T_2 = t_2(x_1, x_2, ..., x_n)$ be a two statistic satisfying the condition of $T_1 \leq T_2$ for which $P_\theta\left[T_1 < \tau(\theta) < T_2\right] \equiv \gamma$ where γ does not depend on θ, then the random interval part $\tau(\theta)$, γ is called confidence coefficient and T_1 *and* T_2 are called lower and upper confidence limits respectively for $\tau(\theta)$. A value t_1, t_2 of the random interval T_1 *and* T_2 is also called a 100γ % confidence interval for $\tau(\theta)$.

Definition 2: ONE SIDED CONFIDENCE INTERVAL

Let $x_1, x_2, ..., x_n$ be a random sample from the density $f(., \theta)$. Let $T_1 = t_1(x_1, x_2, ..., x_n)$be a statistic for which $P_\theta\left[T_1 < \tau(\theta)\right] \equiv \gamma$ then T_1 is called a one sided lower confidence for $\tau(\theta)$. Similarly,

$T_2 = t_2\left(x_1, x_2, \ldots, x_n\right)$be a statistic for which $P_\theta\left[\tau(\theta) < T_2\right] \equiv \gamma$ then T_2 is called a one sided upper confidence for $\tau(\theta)$.

CONSTRUCTION OF CONFIDENCE INTERVAL FOR POPULATION MEAN (when the variance is known)

Let $x_1, x_2, \ldots, x_n$ be a random sample from the normal population with mean μ and variance σ^2. We take a large sample from a normal population with mean μ and SD σ . Then

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

To claim, $100(1-\alpha)\%$ confidence interval for the level of significance at 5% from the normal probability table

$$P[-1.96 \le Z \le 1.96] = 0.95$$

$$\Rightarrow P\left[-1.96 \le \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \le 1.96\right] = 0.95$$

$$\Rightarrow P\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le 1.96\frac{\sigma}{\sqrt{n}} + \bar{x}\right] = 0.95$$

$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ are 95% confidence limit for the unknown parameter μ and the interval $\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \ , \ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is called the 95% confidence interval for μ. Also to construct $100(1-\alpha)\%$ confidence interval for the level of significance at 1% from the normal probability table

$$P[-2.58 \le Z \le 2.58] = 0.99$$

$$\Rightarrow P\left[-2.58 \le \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \le 2.58\right] = 0.99$$

$$\Rightarrow P\left[\bar{x} - 2.58\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}\right] = 0.99$$

$\bar{x} \pm 2.58\frac{\sigma}{\sqrt{n}}$ are 99% confidence limit for the unknown parameter μ and the interval $\left(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}} \ , \ \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}\right)$ is called the 95% confidence interval for μ.

In general, $P\left(-z_\alpha \le z \le z_\alpha\right) = 1 - \alpha$

$$\Rightarrow P\left[-z_\alpha \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_\alpha\right] = 1-\alpha$$

$$\Rightarrow P\left[\bar{x}-z_\alpha\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+z_\alpha\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

Hence the confidence interval for μ is $\left(\bar{x}-z_\alpha\frac{\sigma}{\sqrt{n}}\ ,\ \bar{x}+z_\alpha\frac{\sigma}{\sqrt{n}}\right)$ where z_α is the standard normal value for given level of α .

CONFIDENCE INTERVAL FOR POPULATION MEAN (when variance is unknown)

Let $x_1, x_2, \ldots, x_n$ be a random sample from the normal population with mean μ and variance σ^2. We know that population variance $s^2 = \frac{1}{n-1}\Sigma(x_i-\bar{x})^2$. A statistic $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{(n-1)}$. Hence $100(1-\alpha)\%$ confidence limit for μ is given by

$$P(|t| \le t_\alpha) = 1-\alpha$$

$$\Rightarrow P\left[\left|\frac{\bar{x}-\mu}{s/\sqrt{n}}\right|\right] = 1-\alpha$$

$$\Rightarrow P\left[\bar{x}-t_\alpha\left(\frac{s}{\sqrt{n}}\right) \le \mu \le \bar{x}+t_\alpha\left(\frac{s}{\sqrt{n}}\right)\right] = 1-\alpha$$

where t_α is a tabulated value of student t for (n-1) degrees of freedom at significance level α . Hence required confidence interval for population mean μ is $\left(\bar{x}-t_\alpha\left(\frac{s}{\sqrt{n}}\right)\ ,\ \bar{x}+t_\alpha\left(\frac{s}{\sqrt{n}}\right)\right)$.

CONSTRUCTION OF CONFIDENCE INTERVAL FOR POPULATION VARIANCE (when mean is known)

Let $x_1, x_2, \ldots, x_n$ be a random sample from the normal population with mean μ and variance σ^2. The statistic

$$\frac{\Sigma(x_i-\mu)^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2_{(n)}$$

where $s^2 = \frac{1}{n}\Sigma(x_i-\mu)^2$

Let $\chi^2{}_\alpha$ at the value of χ^2 such that

$$P\left[\chi^2 > \chi^2{}_{\alpha}\right] = \int_{\chi^2{}_{\alpha}}^{\infty} P\left(\chi^2\right) d\chi^2$$

where $P\left(\chi^2\right)$ is the probability density function of χ^2 distribution with n degrees of freedom and significance level α. Thus the required confidence interval is given by

$$P\left[\chi^2{}_{1-\alpha/2} \le \chi^2 \le \chi^2{}_{\alpha/2}\right] = 1-\alpha$$

$$\Rightarrow P\left[\chi^2{}_{1-\alpha/2} \le \frac{ns^2}{\sigma^2} \le \chi^2{}_{\alpha/2}\right] = 1-\alpha$$

Now , $\dfrac{ns^2}{\sigma^2} \le \chi^2{}_{\alpha/2} \qquad \Rightarrow \dfrac{ns^2}{\chi^2{}_{\alpha/2}} \le \sigma^2$

$$\chi^2{}_{1-\alpha/2} \ge \frac{ns^2}{\sigma^2} \qquad \Rightarrow \frac{ns^2}{\chi^2{}_{1-\alpha/2}} \ge \sigma^2$$

Then, $P\left[\dfrac{ns^2}{\chi^2{}_{\alpha/2}} \le \sigma^2 \le \dfrac{ns^2}{\chi^2{}_{1-\alpha/2}}\right] = 1-\alpha$

where $\chi^2{}_{\alpha/2}$ and $\chi^2{}_{1-\alpha/2}$ are obtained from χ^2 table with n degrees of freedom and significant level α.

CONSTRUCTION OF CONFIDENCE INTERVAL FOR POPULATION VARIANCE (When Mean is Unknown)

Let $x_1, x_2, \ldots, x_n$ be a random sample from the normal population with mean μ and variance σ^2.

Here the statistic $\dfrac{\Sigma\left(x_i - \bar{x}\right)}{\sigma^2} = \dfrac{ns^2}{\sigma^2} \sim \chi^2{}_{(n-1)}$

where $s^2 = \dfrac{1}{n}\Sigma\left(x_i - \bar{x}\right)^2$

Let $\chi^2{}_{\alpha}$ as the value of χ^2 such that

$$P\left[\chi^2 > \chi^2{}_{\alpha}\right] = \int_{\chi^2{}_{\alpha}}^{\infty} P\left(\chi^2\right) d\chi^2$$

where $P\left(\chi^2\right)$ is the probability density function with (n-1) degrees of freedom and significance level α .Thus the required confidence interval is given by

$$P\left[\chi^2{}_{1-\alpha/2} \le \chi^2 \le \chi^2{}_{\alpha/2}\right] = 1-\alpha$$

$$\Rightarrow P\left[\chi^2{}_{1-\alpha/2} \le \frac{ns^2}{\sigma^2} \le \chi^2{}_{\alpha/2}\right] = 1-\alpha$$

$$\Rightarrow P\left[\frac{1}{\chi^2_{1-\alpha/2}} \le \frac{\sigma^2}{ns^2} \le \frac{1}{\chi^2_{\alpha/2}}\right] = 1-\alpha$$

$$\Rightarrow P\left[\frac{ns^2}{\chi^2_{1-\alpha/2}} \ge \sigma^2 \ge \frac{ns^2}{\chi^2_{\alpha/2}}\right] = 1-\alpha$$

$$\Rightarrow P\left[\frac{ns^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{ns^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are obtained from χ^2 table with (n-1) degrees of freedom and significant level α.

CONSTRUCTION OF CONFIDENCE INTERVAL FOR DIFFERENCE OF MEANS OF TWO INDEPENDENT NORMAL POPULATION WHEN VARIANCE IS KNOWN

Let $x_1, x_2, \ldots, x_n \sim N(\mu_x, \sigma^2{}_x)$ and $y_1, y_2, \ldots, y_n \sim N(\mu_y, \sigma^2{}_y)$. The statistic

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}}$$

The required confidence interval for given level of significance

$$P[-z_{\alpha/2} \le z \le z_{\alpha/2}] = 1-\alpha$$

$$\Rightarrow P\left[-z_{\alpha/2} \le \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}} \le z_{\alpha/2}\right] = 1-\alpha$$

$$\Rightarrow P\left[-z_{\alpha/2}\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}} \le \bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2) \le z_{\alpha/2}\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}\right] = 1-\alpha$$

$$\Rightarrow P\left[(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}} \le (\mu_1 - \mu_2) \le (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}\right] = 1-\alpha$$

Hence the difference of population mean confidence interval for the given level of significance α is given by $\left((\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}\ ,\ (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}\right)$ and the confidence limit is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}$$

CONSTRUCTION OF CONFIDENCE INTERVAL FOR DIFFERENCE OF MEANS OF TWO INDEPENDENT NORMAL POPULATION WHEN VARIANCE IS UNKNOWN

$$x_1, x_2, \ldots, x_n \sim N(\mu_x, s^2{}_x) \qquad y_1, y_2, \ldots, y_n \sim N(\mu_y, s^2{}_y)$$

The statistic $z = \dfrac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s^2{}_1}{n_1} + \dfrac{s^2{}_2}{n_2}}}$

The required confidence interval for given level of significance

$$P[-z_{\alpha/2} \le z \le z_{\alpha/2}] = 1 - \alpha$$

$$\Rightarrow P\left[-z_{\alpha/2} \le \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}} \le z_{\alpha/2}\right] = 1 - \alpha$$

$$\Rightarrow P\left[-z_{\alpha/2}\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}} \le \bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2) \le z_{\alpha/2}\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}\right] = 1 - \alpha$$

$$\Rightarrow P\left[(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}} \le (\mu_1 - \mu_2) \le (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}\right] = 1 - \alpha$$

Hence the difference of population mean confidence interval for the given level of significance α is given by $\left((\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}\ ,\ (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}\right)$ and the confidence limit is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}$$

CONSTRUCTION OF CONFIDENCE INTERVAL FOR DIFFERENCE OF MEANS OF TWO SAMPLES OF NORMAL POPULATION WITH COMMON VARIANCE (COMMON VARIANCE IS UNKNOWN)

Let $x_1, x_2, \ldots, x_m$ be a random sample from the normal population with mean μ_1 and variance $\sigma^2{}_1$. Let $y_1, y_2, \ldots, y_n$ be a random sample from the normal population with mean μ_2 and variance $\sigma^2{}_2$. Assume that the two samples are independent to each other. Let $\bar{y} - \bar{x}$ is normally distributed with mean $\mu_2 - \mu_1$ and variance $\frac{\sigma_2}{m} + \frac{\sigma^2}{n}$ (i.e) $(\bar{y} - \bar{x}) \sim N\left(\mu_2 - \mu_1 \ , \frac{\sigma_2}{m} + \frac{\sigma^2}{n}\right)$. $\frac{\Sigma(x_i - \bar{x})^2}{\sigma^2}$ is chi-square distributed with (m-1) degrees of freedom

$$\frac{\Sigma(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2{}_{(m-1)} \quad \text{and}$$

$$\frac{\Sigma(y_i - \bar{y})^2}{\sigma^2} \sim \chi^2{}_{(n-1)}$$

$$\therefore \frac{\Sigma(x_i - \bar{x})^2}{\sigma^2} + \frac{\Sigma(y_i - \bar{y})^2}{\sigma^2} \sim \chi^2{}_{(m+n-2)}$$

The statistic

$$Q = \frac{(\bar{y} - \bar{x}) - (\mu_2 - \mu_1)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right). s_{p^2}}} \sim F_{(m+n-2)}$$

Thus the confidence interval for difference of means for two samples of normal population with the given level of significance α

$$P[-t_{\alpha/2} \le Q \le t_{\alpha/2}] = 1 - \alpha$$

$$\Rightarrow P\left[-t_{\alpha/2} < \frac{(\bar{y} - \bar{x}) - (\mu_2 - \mu_1)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right). s_{p^2}}} < t_{\alpha/2}\right] = 1 - \alpha$$

$$\Rightarrow P\left[(\bar{y} - \bar{x}) - t_{\alpha/2}\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right). s_{p^2}} < (\mu_2 - \mu_1) < (\bar{y} - \bar{x}) + t_{\alpha/2}\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right). s_{p^2}}\right] = 1 - \alpha$$

Hence $100(1 - \alpha)\%$ confidence interval is $\left((\bar{y} - \bar{x}) - t_{\alpha/2}\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right). s_{p^2}} \ , \ (\bar{y} - \bar{x}) + t_{\alpha/2}\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right). s_{p^2}}\right)$ and the confidence limits are

$$(\bar{y} - \bar{x}) \pm t_{\alpha/2}\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right). s_{p^2}}$$

Suppose the samples are dependent on each other with common variance. Let $D_i = y_i - x_i \quad \forall\ i = 1,2,\ldots,n \text{ then } D_1, D_2, \ldots, D_n$ are independently identically distributed random variables with common normal distribution having mean $\mu_D = \mu_2 - \mu_1$ and variance $\sigma^2{}_D = \sigma_1{}^2 + \sigma_2{}^2 - 2\rho\sigma_1\sigma_2$

$\therefore 100\left(1-\alpha\right)\%$ confidence interval for $\mu_D = \mu_2 - \mu_1$ is

$\left(\overline{D} - t_{\alpha/2}\sqrt{\frac{\Sigma\left(D_i - \overline{D}\right)^2}{n\left(n-1\right)}}\ ,\ \overline{D} + t_{\alpha/2}\sqrt{\frac{\Sigma\left(D_i - \overline{D}\right)^2}{n\left(n-1\right)}}\right)$ where $t_{\alpha/2}$ is the $\alpha/2$ th quartile point of the t-distribution with (n-1) degrees of freedom.

UNIT-III

In the previous chapter, we have discussed methods of point estimation which are expected to yield estimators enjoying some of these important properties. Also we have discussed the confidence interval for proportions, mean(s), variance(s) based on chi-square, Student's t, F and Normal Distributions. In this chapter, we shall discuss the statistical hypothesis. A statistical hypothesis is some statement or assertion about a population or equivalently about the probability distribution characterising a population which we want to verify on the basis of information available from a sample.

Simple and Composite Hypothesis:

When a hypothesis specifies all the parameters of a probability distribution, it is known as simple hypothesis. The hypothesis specifies all the parameters, i.e μ and σ of a normal distribution.

Example: The random variable x is distributed normally with mean $\mu=0$ & SD=1 is a simple hypothesis. The hypothesis specifies all the parameters (μ & σ) of a normal distributions.

If the hypothesis specific only some of the parameters of the probability distribution, it is known as composite hypothesis. In the above example if only the μ is specified or only the σ is specified it is a composite hypothesis.

Test of Statistical Hypothesis:

A test of statistical hypothesis is a two action decision problem after the experimental sample value has been obtained. The two action being acceptance rejection of the hypothesis under consideration.

Null Hypothesis: In hypothesis, testing a decision maker should not be motivated by prospects of profit or loss resulting from the acceptance or rejection of the hypothesis, ie., neutral or general statement about the population parameter is known as null hypothesis.

Alternative Hypothesis: it is desirable to reject the hypothesis based on statistical test in other words, the general statement which is opposite to be null hypothesis stated is known as alternative hypothesis.

Critical Region: let $x_1, x_2, ..., x_n$ be the sample observation denoted by 0. We specify some region of the n dimensional space and see whether this point lies within this region or outside this region. We divide the whole sample space into two disjoint regions $w \; and \; \overline{w}(s-w)$.

The null hypothesis H_0 is rejected if the observed sample point falls in $\overline{w}$ and if it falls in we accept H_0 i.e the region of rejection of H_0 when H_0 is true is that region of the outcome set where H_0 is rejected. If the sample point falls in that region then it is called critical region.

Type I Error: rejecting the null hypothesis H_0 when is true is called type I error.

Type II Error: the error of accepting H_0 when it false is called type II error.

Level of Significance: probability of type I error is known as level of significance of test. It is also called as size of the critical region.

$$\alpha = p[type\, I\, error]$$

$$\alpha = p[x\,\varepsilon\, w / H_0]$$

$$\alpha = \int_w L_0 \, dx$$

where L_0 is the likelihood function of the sample observation under H_0.

Power of the Test: probability of type II error is denoted by β . $1-\beta$ is called power function of the hypothesis against the alternative H_1 . The value of the power function at a parameter point is called power of the test at that point (i.e).

$$\beta = p[type\, II\, error]$$

$$\beta = p[x\,\varepsilon\, w / H_1]$$

$$\beta = \int_{\overline{w}} L_1 \, dx$$

We have,

$$\int_w L_1 \, dx + \int_{\overline{w}} L_1 \, dx = 1$$

$$\int_w L_1 \, dx + \beta = 1$$

$$\int_w L_1 \, dx = 1 - \beta$$

STEPS INVOLVED IN TESTING OF HYPOTHESIS:

- Explicit knowledge of the nature of population distribution and the parameter of interest (i.e) the parameter about which the hypothesis are setup
- Setting up the null hypothesis H_0 and the alternative hypothesis H_1 in terms of the range of parameter values each ones embodies.
- The choice of a suitable statistic called the test statistic which will be reflecting upon the probability of H_0 and H_1.
- Partitioning the set of possible values of the test statistic into two disjoint sets w and $\overline{w}$ and framing the following test.
    - Reject H_0 if the value of test statistic falls in w (critical region)
    - Accept H_0 if falls in $\overline{w}$ (acceptance region)
- After framing the above obtain experimental sample observation, compute the appropriate test statistic and take actions accordingly.

Example 1: A single observation is taken from Poisson population to test $H_0 : \lambda = 2$ against $H_1 : \lambda = 3$ based on the critical region $w = \{x : x \geq 4\}$ .find α, β and power of the test.

Solution:

The probability distribution of population is given by,

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!} \qquad ; x = 0,1,\ldots,\lambda > 0$$

Given that $H_0 : \lambda = 2$

$H_1 : \lambda = 3$

Critical region: $w = \{x : x \geq 4\}$

Acceptance region: $\overline{w} = \{x : x < 4\}$

$\alpha = p[type\, I\, error]$

$\alpha = p[x \, \varepsilon \, w / H_0]$

$$=\sum_{x=4}^{\infty}\frac{e^{-2}2^{x}}{x!}$$

$$=\sum_{x=0}^{\infty}\frac{e^{-2}2^{x}}{x!}-\sum_{x=3}^{3}\frac{e^{-2}2^{x}}{x!}$$

$$=1-\sum_{x=3}^{3}\frac{e^{-2}2^{x}}{x!}$$

$$=1-e^{-2}\left[\frac{2^{0}}{0!}+\frac{2^{1}}{1!}+\frac{2^{2}}{2!}+\frac{2^{3}}{3!}\right]$$

$$=1-e^{-2}\left[1+2+2+\frac{8}{6}\right]$$

$$=1-0.1353\left(\frac{19}{3}\right)$$

$$\alpha=0.1431$$

$$\beta=p[type\,II\,error]$$

$$\beta=p[x\,\varepsilon\,\overline{w}/H_1]$$

$$=\sum_{x=0}^{3}\frac{e^{-3}3^{x}}{x!}$$

$$=e^{-3}\left[\frac{3^{0}}{0!}+\frac{3^{1}}{1!}+\frac{3^{2}}{2!}+\frac{3^{3}}{3!}\right]$$

$$=0.0498\left[1+3+\frac{9}{2}+\frac{27}{6}\right]$$

$$=0.0498*13$$

. $\beta=0.6474$

Power of the test=$1-\beta$

=1-0.6474=0.3256

Example 2: A single observation is taken from binomial population to test $H_0: p=1/2$ against $H_1: p=3/4$ based on the critical region $w=\{x: x>4\}$ where x denotes the number of heads when the coin is tossed 6 times.

Solution:

The probability mass function of binomial distribution is given by

$$p(x) = \binom{n}{x} p^x q^{n-x} \quad x = 0,1,\ldots,n \quad ; p + q = 1$$

Given that $H_0 : p = 1/2$

$H_1 : p = 3/4$

Critical region: $w = \{x : x < 4\}$

Acceptance region: $\overline{w} = \{x : x \geq 4\}$

$\alpha = p[type\,I\,error] = p[x\,\varepsilon\,w / H_0]$

$$= \sum_{x=5}^{6} \binom{6}{x} p^x q^{6-x}$$

$$= \sum_{x=5}^{6} \binom{6}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{6-x}$$

$$= 6C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^1 + 6C_6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^0 = 0.0938 + 0.0156$$

$\alpha = 0.1094$

$\beta = p[type\,II\,error] = p[x\,\varepsilon\,\overline{w} / H_1]$

$$= \sum_{x=0}^{4} \binom{6}{x} \left(\frac{3}{4}\right)^x \left(\frac{1}{4}\right)^{6-x}$$

$$= 6C_0 \left(\frac{3}{4}\right)^0 \left(\frac{1}{4}\right)^6 + 6C_1 \left(\frac{3}{4}\right)^1 \left(\frac{1}{4}\right)^5 + 6C_2 \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^4 + 6C_3 \left(\frac{3}{4}\right)^3 \left(\frac{1}{4}\right)^3 + 6C_4 \left(\frac{3}{4}\right)^4 \left(\frac{1}{4}\right)^2$$

$$= 1 * \frac{1}{4096} + 6 * \frac{3}{4} \frac{1}{1024} + 15 * \frac{9}{16} \frac{1}{256} + 20 * \frac{27}{64} \frac{1}{64} + 15 * \frac{81}{256} \frac{1}{16}$$

=0.00024+ 0.00439 + 0.03296 + 0.13184 + 0.29663 =0.4661

Power of the test = $1 - \beta$=1-0.4661=0.5339

Example 3: A single observation is taken from exponential family to test $H_0 : \theta = 2$ against $H_1 : \theta = 1$ and agreed to reject H_0 when $x \geq 1$. Find α, β and power of the test.

Solution:

The probability density function of exponential distribution is given by,

$$f(x) = \theta e^{-\theta x} \quad ;\theta > 0\ ,\ \ 0 < x < \infty$$

Given that $H_0 : \theta = 2$

$H_1 : \theta = 1$

Critical region: $w = \{x : x \geq 1\}$

Acceptance region: $\overline{w} = \{x : x < 1\}$

$$\alpha = p[type\,I\,error] = p[x\,\varepsilon\,w / H_0]$$

$$= \int_1^\infty \theta e^{-\theta x} dx$$

$$= \theta\left[-\frac{1}{\theta}e^{-\theta x}\right]_1^\infty = \left[-e^{-2x}\right]_1^\infty$$

$$\alpha = -e^{-\infty} + e^{-2}$$

=0.1353

$$\beta = p[type\,II\,error] = p[x\,\varepsilon\,\overline{w} / H_1]$$

$$= \int_1^\infty \theta e^{-\theta x} dx$$

$$= 1\left[-\frac{1}{1}e^{-x}\right]_0^1$$

$$\beta = -e^{-1} + e^{-0}$$

$$\beta = 0.6321$$

Power of the test = $1 - \beta$

=1-0.6321=0.3679

Example 4: A single observation is taken from the probability distribution $f(x,\theta) = \frac{1}{\theta}$ $0 \leq x \leq \theta\ ;\theta > 0$ to test $H_0 : \theta = 1$ against $H_1 : \theta = 2$ and agreed to reject H_0 when $x \geq 0.5$. Find α, β and power of the test.

Solution:

The probability density function of uniform distribution is given by

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

Given that $H_0 : \theta = 1$, $H_1 : \theta = 2$

Critical region: $w = \{x : x \geq 0.5\}$

Acceptance region: $\overline{w} = \{x : x < 0.5\}$

$$\alpha = p[type\, I\, error] = p[x \,\varepsilon\, w / H_0]$$

$$\alpha = p[x \geq 0.5 / \theta = 1]$$

$$= \int_{0.5}^{1} \frac{1}{\theta}$$

. $$\alpha = [x]_{0.5}^{1} = 1 - 0.5 = 0.5$$

$$\beta = p[type\, II\, error] = p[x \,\varepsilon\, \overline{w} / H_1]$$

$$= \int_{0.5}^{1} \frac{1}{\theta}$$

$$\beta = \frac{1}{2}[x]_{0.5}^{1} = \frac{0.5}{2} = 0.25$$

Power of the test = $1 - \beta$

=1-0.25=0.75

Example 5: $f(x) = (1+\theta)x^{\theta}$; $0 \leq x \leq 1$. A single observation is taken from the given distribution. Find α, β and power of the test if the test is $H_0 : \theta = 1$ against $H_1 : \theta = 2$ based on the critical region when $x \leq 0.5$.

Solution:

Given that $H_0 : \theta = 1$; $H_1 : \theta = 2$

Critical region: $w = \{x : x \leq 0.5\}$

$$\alpha = p[type\, I\, error] = p[x \,\varepsilon\, w / H_0]$$

$$= \int_{0}^{0.5} (1+\theta)x^{\theta}\, dx$$

$$= \int_{0}^{0.5} (1+1)x^{1} dx$$

$$\alpha = 2\left[\frac{x^2}{2}\right]_{0}^{0.5}$$

$$\alpha = 0 + (0.5)^2$$

$$\alpha = 0.25$$

$$\beta = p[type\, II\, error] = p[x \,\varepsilon\, \overline{w} / H_1]$$

$=\int_{0.5}^{1}(1+2)x^2dx$ Acceptance region: $\overline{w}=\{x:x>0.5\}$

$=1-0.125$

$\beta=0.875$

Power of the test = $1-\beta$

=1-0.875=0.125

Example 6: A single observation is taken from the $f(x,\theta)=\theta e^{-\vartheta x}$ $;\theta>0;0\leq x\leq\infty$ to test $H_0:\theta=2$ against $H_1:\theta=1$. Find the best critical region of single 0.05.

Solution:

Given that $f(x,\theta)=\theta e^{-\vartheta x}$

$$H_0:\theta=2$$

$$H_1:\theta=1$$

$$\alpha=0.05$$

$\alpha=p[type\,I\,error]$

Let critical region = $\{x:x\geq x_0\}$ where x_0 is a constant which is to be estimated let the probability density function under H_0 is given by

$$f(x,\theta_0)=2e^{-2x}\,;x>0$$

Similarly the probability density function under H_1 is given by

$$f(x,\theta_1)=e^{-x}\,;x>0$$

$$\alpha=p[type\,I\,error]=p[x:x\geq x_0/H_0]$$

$$0.05=\int_{x_0}^{\infty}f(x,\theta_0)$$

$$=\int_{x_0}^{\infty}2e^{-2x}dx$$

$$=2\left[\frac{-1}{2}e^{-2x}\right]_{x_0}^{\infty}$$

$$=-e^{-\infty}+e^{-2x_0}$$

$$=e^{-2x_0}$$

Taking log on both sides,

`log 0.05=-2 x_0

-2.9957=-2 x_0

x_0 =1.4979

The required critical region of sixe 0.05 is $w = \{x : x \geq 1.4979\} \cong w = \{x : x \geq 1.5\}$

Acceptance region: $\overline{w} = \{x : x < 1.5\}$

$\beta = p[type\, II\, error] = p[x\, \varepsilon\, \overline{w} / H_1]$

$= \int_0^{1.5} e^{-x} dx$

$= \left[-e^{-x}\right]_0^{1.5}$

$= -e^{-1.5} + e^0$

=-0.2331+1

$\beta = 0.7769$

Power of the test = $1-\beta$

=1-0.7769=0.2231

Case II:

Critical region: $w = \{x : x \leq x_0\}$

$\alpha = p[type\, I\, error] = p[x : x \leq x_0 / H_0]$

$\int_0^{x_0} 2e^{-2x} dx = 0.05$

$2\left[\frac{-1}{2} e^{-2x}\right]_0^{x_0} = 0.05$

$e^0 - e^{-2x_0}$ =0.05

$-e^{-2x_0}$ $=-$ 0.95

e^{-2x_0} =0.95

Taking log on both sides,

log 0.95=-2 x_0

-0.0513=-2 x_0

x_0 = 0.0257

The required critical region of sixe 0.05 is $w = \{x : x \leq 0.0257\} \cong w = \{x : x \leq 0.026\}$

Acceptance region: $\overline{w} = \{x : x > 0.026\}$

$\beta = p[type\, II\, error] = p[x : x\,\varepsilon\, \overline{w} / H_1]$

$$= \int_{0.026}^{\infty} e^{-x} dx$$

$$= \left[-e^{-x}\right]_{0.026}^{\infty}$$

$$= -e^{-\infty} + e^{-0.026}$$

$$\beta = 0 + 0.9743$$

$$\beta = 0.9743$$

Power of the test = $1 - \beta$

=1-0.9743=0.0257

In the given example we have possible critical region of size 0.05

Size	Critical region	Power of test
0.05	$w = \{x : x \geq 1.5\}$	0.2231
0.05	$w = \{x : x \leq 0.026\}$	0.0257

We select the first critical region because it has maximum power of the test.

Randomized Tests

It will be recalled that for hypothesis testing problems involving *discrete* distributions, it is usually not possible to choose a critical region consisting of realizable values of the statistic of size exactly α, where α is some prescribed value.

In the hypothesis testing procedures considered so far, the sample space of observations X is partitioned into 2 regions, C and $\overline{C}$ (its complement). We can express this in terms of a function ψ as follows. Let

$$\psi(x) = P(reject\ \ H_0\ \ when\ \ X = x)$$

For a non-randomized test with rejection region C, ψ for a region C is just its indicator function. That is,

$$\psi(x) = \begin{cases} 1 & if \quad x \in C \\ 0 & if \quad x \notin C \end{cases}$$

We will extend this, to allow for some different action (other that ``reject'' and ``accept'') if the outcome **x** is on the boundary of the critical region. The other action effectively is performing an auxiliary experiment such as tossing a coin with P(heads) $=$p; if heads results, reject H$_0$; if tails results, H$_0$ is accepted. The value of p is chosen to make the P(rejecting H$_0$) the desired value. More formally, for a test with critical region C and a value of X= x$_0$ on the boundary, we may define

$$\psi(x)=\begin{cases}1 & if \quad x\in C\\ p & if \quad x=x_0\\ 0 & if \quad x\neq x_0 \quad and \quad x\notin C\end{cases}$$

where p (0<p<1) is appropriately chosen.

Best Critical Region and Most Powerful Test:

A critical region w of size α for testing H_0 against H_1 is said to be best critical region (BCR) if w^* is any other critical of same size α for which power of $w\geq$ power of w^* (i.e).

$$1\text{- } p[x\,\varepsilon\,\overline{w}/H_1]\geq 1-p[x\,\varepsilon\,\overline{w}^*/H_1] \text{ or}$$
$$p[x\,\varepsilon\,\overline{w}/H_1]\geq p[x\,\varepsilon\,\overline{w}^*/H_1]$$

A statistical test based on best critical region is called most powerful test.

Neymann Pearson's Fundamental Lemma:

Let $x_1,x_2,...,x_n$ be a random sample from $f(x,\theta)$where θ is the unknown parameter. Let L_0 and L_1 be the likelihood functions under $H_0:\theta=\theta_0$ against $H_1:\theta=\theta_1$ respectively if there exist a critical region w of size α at a constant k such that $\frac{L_0}{L_1}\leq k$ for points in$w$. Then w is the best critical region of size for testing H_0 against H_1.

Proof:

Let w^* be any other critical region of sizeα .

$$\alpha = p[type\,I\,error] = \text{size of w}$$
$$=\int_w L_0 dx$$

$\alpha = size\ of\ w^*$

$= \int\limits_{w^*} L_0 dx$

$\alpha = \int\limits_{w} L_0 dx = \int\limits_{w^*} L_0 dx$

$\alpha = \int\limits_{a} L_0 dx = \int\limits_{c} L_0 dx = 0$ ---------------(1)

Power of w $= 1 - \beta$

$= 1 - \int\limits_{\overline{w}} L_1 dx$

$= \int\limits_{\overline{w}} L_1 dx$

Power of $\overline{w}$ $= 1 - \beta$

$= 1 - \int\limits_{\overline{w}^*} L_1 dx$

$= \int\limits_{\overline{w}^*} L_1 dx$

Power of w - Power of $\overline{w}$ $= \int\limits_{\overline{w}} L_1 dx$ - $\int\limits_{\overline{w}^*} L_1 dx$ --------------(2)

From the lemma $a\ \varepsilon\ w$ and $\frac{L_0}{L_1} \le k$

$L_0 \le k\ \ L_1$

$\int\limits_{a} L_0 dx\ \le k\ \int\limits_{a} L_1 dx$

$\int\limits_{a} L_0 dx\ \ge \frac{1}{k} \int\limits_{a} L_1 dx$ -----------------(3)

Conversely,

$c\ \varepsilon\ w$ and $\frac{L_0}{L_1} \ge k$

$L_0 \ge k\ \ L_1$

$L_1\ \le \frac{1}{k}\ L_0$

$\int\limits_{c} L_1 dx\ \ \le \frac{1}{k} \int\limits_{c} L_0 dx$

$$-\int_c L_1 dx \ \geq \frac{1}{k} \int_c L_0 dx \qquad \text{---------------(4)}$$

From (2)

Power of w - Power of w^* = $\int_{\bar{w}} L_1 dx$ - $\int_{\bar{w}^*} L_1 dx$

$$= \int_a L_1 dx \ - \int_c L_1 dx$$

$$\geq \frac{1}{k} \int_a L_0 dx \ + \left(-\frac{1}{k}\right) \int_c L_0 dx \qquad \text{[using (3) \& (4)]}$$

$$\geq \frac{1}{k} \left[\int_a L_0 dx \ - \int_c L_0 dx \right]$$

$$\geq 0$$

Power of w - Power of w^* ≥ 0

Power of w $\geq$ Power of w^*

w is a best critical region.

Definition 1: A critical region w of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is said to be uniformly most powerful critical region if for every value of $\theta \neq \theta_0$ the power of the critical region w must be greater than or equal to the critical region w must be greater than or equal to power of any other critical region w^* of same size α any test based on uniformly most powerful critical region is called **uniformly most powerful test**.

Example 7: Given a random sample $x_1, x_2, ..., x_n$ from the distribution with the pdf $f(x,\theta) = \theta e^{-\theta x}$; $x > 0$; $\theta > 0$ show that there exist no UMPT for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

Solution:

Let $x_1, x_2, ..., x_n$ be a random sample from exponential distribution then the likelihood function is given by

$$f(x,\theta) = \theta e^{-\theta x}$$

$$\frac{L_0}{L_1} \le k$$

$$\prod_{i=1}^{n} f\left(x_i, \theta_0\right) = \prod_{i=1}^{n} \theta_0 e^{-\theta_0 x_i}$$

$$\frac{L_0}{L_1} = \frac{\theta_0{}^{n} e^{-\theta_0 \sum_{i=1}^{n} x_i}}{\theta_1{}^{n} e^{-\theta_1 \sum_{i=1}^{n} x_i}} \le k$$

$$\left(\frac{\theta_0}{\theta_1}\right)^{n} e^{-\sum_{i=1}^{n} x_i \left(\theta_0 - \theta_1\right)} \le k$$

Taking log on both sides,

$$n \log\left(\frac{\theta_0}{\theta_1}\right) \; - \sum_{i=1}^{n} x_i \left(\theta_0 - \theta_1\right) \le \log k$$

$$n\left[\log \theta_0 - \log \theta_1\right] - \sum_{i=1}^{n} x_i \left(\theta_0 - \theta_1\right) \le \log k$$

$$-\sum_{i=1}^{n} x_i \left(\theta_0 - \theta_1\right) \le \log k - n\left[\log \theta_0 - \log \theta_1\right]$$

Case I: $\theta_0 > \theta_1$

$$\Rightarrow \theta_0 - \theta_1 > 0$$

$\log \theta_0 - \log \theta_1$ is a positive quantity

$$\sum_{i=1}^{n} x_i \ge \frac{\log k - n\left[\log \theta_0 - \log \theta_1\right]}{\theta_0 - \theta_1}$$

Case II: $\theta_0 < \theta_1$

$\log \theta_0 - \log \theta_1$ is a positive quantity

$$\sum_{i=1}^{n} x_i \le \frac{\log k + n\left[\log \theta_0 - \log \theta_1\right]}{\theta_0 - \theta_1}$$

Case I: $\theta_1 > \theta_0$ then the BCR is given by

$$\sum x_i \le \frac{k_1}{\theta_0 - \theta_1} = \lambda_1 \, (say)$$

Case II: $\theta_0 > \theta_1$ then the BCR is given by

$$\sum x_i \ge \frac{k_1}{\theta_0 - \theta_1} = \lambda_2 \, (say)$$

The constant λ_1 and λ_2 are determined such that

$$p\left[\sum x_i \leq \lambda_1 / H_0\right] = \alpha$$

$$p\left[\sum x_i \geq \lambda_2 / H_0\right] = \alpha$$

Note that if $x \sim E(\theta)$ then $2\theta \sum x_i \sim \chi^2_{2h}$

$$p\left(2\theta \sum x_i\right) = p\left[2\theta \sum x_i \leq 2\theta\lambda_1 / H_0\right] = \alpha$$

$$p\left(2\theta \sum x_i\right) = p\left[2\theta \sum x_i \geq 2\theta\lambda_2 / H_0\right] = \alpha$$

Using this result,

$$p\left[2\theta \sum x_i \leq \mu_1\right] = p\left[\chi^2_{2h} \leq \mu_1\right] = \alpha$$

$$\chi^2_{2h} = \mu_1$$

Hence the BCR for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 \ (> \theta_0)$ is given by

$$w_0 = \left\{x_i : 2\theta \sum x_i \leq \chi^2_{1-\alpha, 2n}\right\}$$

$$w_0 = \left\{x_i : \sum x_i \leq \chi^2_{1-\alpha, 2n/2\theta}\right\}$$

Since w_0 is independent of w_0, θ_0 is UMPCR for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 \ (> \theta_0)$ similarly

$$p\left[2\theta \sum x_i \geq 2\theta\lambda_2\right]$$

$\alpha = p\left[\chi^2_{2h} \geq \mu_1\right]$ where –

Hence, BCR for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 \ (< \theta_0)$ is given by

$$w_1 = \left\{x_i : 2\theta \sum x_i \geq \chi^2_{1-\alpha, 2n}\right\}$$

$$w_1 = \left\{x_i : \sum x_i \geq \chi^2_{1-\alpha, 2n/2\theta}\right\}$$

Since w_1 is independent of w_1, θ_1 is UMPCR for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 \ (< \theta_0)$ similarly. Since the two CR w_0 and w_0 are different there exists no CR of size α which is UMP for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

Power of the test:

$$1 - \beta = p\left[x \varepsilon w_0 / H_1\right]$$

$$= p\left[\sum x_i \leq \frac{1}{2\theta}, \chi^2_{1-\alpha, 2n} / H_1\right]$$

$$= p\left[2\theta_1 \sum x_i \leq \frac{\theta_1}{\theta_0}, \chi^2_{1-\alpha, 2n} / H_1\right]$$

$$= p\left[\chi^2_{2n} \le \frac{\theta_1}{\theta_0}, \chi^2_{1-\alpha,2n}\right]$$

The power of test H for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 \quad (< \theta_0)$ is given by.

$$1-\beta = p[x\varepsilon w_0 / H_1]$$

$$= p\left[\sum x_i \ge \frac{1}{2\theta}, \chi^2_{1-\alpha,2n} / H_1\right]$$

$$= p\left[2\theta_1 \sum x_i \ge \frac{\theta_1}{\theta_0}, \chi^2_{1-\alpha,2n} / H_1\right]$$

Example 8: show that for a normal distribution with mean 0 and variance – the BCR for testing $H_0 : \sigma = \sigma_0$ versus $H_1 : \sigma = \sigma_1$ is the form $\sum x_i^{\,2} \le a_\alpha$ for $\sigma_0 > \sigma_1$ and $\sum x_i^{\,2} \le b_\alpha$ for $\sigma_0 < \sigma_1$.Show that power of the test of the BCR where $\sigma_0 > \sigma_1$ is $F\left(\frac{\sigma_0^2}{\sigma_1^2}, \chi^2_{\alpha,n}\right)$.

Solution:

Let $x_1, x_2, ..., x_n$ be a random sample of size n from normal population with mean 0 and variance σ^2 .

The likelihood function for $N(0, \sigma_0^2\ )$ is

$$L_0 = \prod_{i=1}^{n} f(x_i, 0, \sigma_0^2\ ) = \left(\frac{1}{\sigma_0^{\,2}\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - 0}{\sigma_0}\right)^2}$$

$$= \left(\frac{1}{\sigma_0^{\,2}\, 2\pi}\right)^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{x_i^2}{\sigma_0^2}}$$

The likelihood function for $N(0, \sigma_1^2\ )$ is

$$L_1 = \prod_{i=1}^{n} f(x_i, 0, \sigma_1^2\ ) = \left(\frac{1}{\sigma_1^{\,2}\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - 0}{\sigma_1}\right)^2}$$

The best critical region is given by

$$\frac{L_0}{L_1} \le k$$

$$\frac{\left(\frac{1}{\sigma_0{}^2 2\pi}\right)^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{x_i^2}{\sigma_0^2}}}{\left(\frac{1}{\sigma_1{}^2 2\pi}\right)^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{x_i^2}{\sigma_1^2}}} \le k$$

$$\left(\frac{\sigma_1^2}{\sigma_0^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\left[\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right]} \le k$$

$$\left(\frac{\sigma_1}{\sigma_0}\right)^{n} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\left[\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right]} \le k$$

Taking log on both sides

$$n\log\left(\frac{\sigma_1}{\sigma_0}\right) \; -\frac{1}{2}\sum_{i=1}^{n} x_i^2\left[\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right] \le \log k$$

$$n[\log\sigma_1 - \log\sigma_0] - \frac{1}{2}\sum_{i=1}^{n} x_i^2\left[\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right] \le \log k$$

$$-\sum_{i=1}^{n} x_i^2\left[\sigma_1^2-\sigma_0^2\right] \le \log k - n[\log\sigma_1 - \log\sigma_0]2\sigma_1^2 * \sigma_0^2$$

$$\sum_{i=1}^{n} x_i^2\left[\sigma_0^2-\sigma_1^2\right] \le [\log k - n\log\sigma_1 + \log\sigma_0]2\sigma_1^2 * \sigma_0^2$$

$$\sum_{i=1}^{n} x_i^2\left[\sigma_0-\sigma_1\right] \le \frac{[\log k - n\log\sigma_1 + \log\sigma_0]2\sigma_1^2 * \sigma_0^2}{\sigma_0+\sigma_1}$$

$$\sum_{i=1}^{n} x_i^2\left[\sigma_0-\sigma_1\right] \le c \quad \text{-----------(1)}$$

Case I: $\sigma_0 < \sigma_1$

Critical region is $\sum_{i=1}^{n} x_i^2\left[\sigma_0-\sigma_1\right] \le c$

$$\sum_{i=1}^{n} x_i^2 \ge \frac{c}{\sigma_0-\sigma_1} \qquad \Rightarrow \sum_{i=1}^{n} x_i^2 \ge c_1$$

Where, $c_1 = \frac{c}{\sigma_0-\sigma_1}$

Case II: $\sigma_0 > \sigma_1$

$$\sum_{i=1}^{n} x_i^2 \leq \frac{c}{\sigma_0 - \sigma_1} \qquad \Rightarrow \sum_{i=1}^{n} x_i^2 \leq c_1$$

Where, $$c_1 = \frac{c}{\sigma_0 - \sigma_1}$$

Case i: $\sigma_0 > \sigma_1$

$$w_1 = \{x_i : \sum x_i \leq a_\alpha\}$$ a_α is determined so that

$$p[x\varepsilon w_1 / H_0] = \alpha$$

$$p\left[\sum_{i=1}^{n} x_i^2 \leq a_\alpha / H_0\right] = \alpha$$

$$p\left[\sum_{i=1}^{n} \frac{x_i^2}{\sigma_0^2} \leq \frac{a_\alpha}{\sigma_0^2} / H_0\right] = \alpha$$

$$p\left[\chi^2 \leq \frac{a_\alpha}{\sigma_0^2}\right] = \alpha \qquad x_i \sim N(0, \sigma_0^2) then\ x_i{}^2 \sim \chi_n^2$$

$$\frac{a_\alpha}{\sigma_0^2} = \chi_{\alpha,2n}^2 \qquad or \qquad a_\alpha = \chi_{\alpha,2n}^2 . \sigma_0^2$$

Hence BCR for testing $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma = \sigma_1 \ (< \sigma_0)$ is given by

$$w = \{x_i : \sum x_i \leq \chi_{\alpha,2n}^2 . \sigma_0^2\}$$

The power of the test is,

$$1 - \beta = p\lfloor x\varepsilon w\ / H_1 \rfloor$$

$$= p\left[\sum_{i=1}^{n} x_i^2 \leq a_\alpha / H_1\right]$$

$$= p\left[\frac{\sum x_i^2}{\sigma_0^2} \leq \frac{a_\alpha}{\sigma_0^2} / H_1\right]$$

$$= p\left[\frac{\sum x_i^2}{\sigma_0^2} \leq \chi_{\alpha,n}^2 / H_1\right]$$

$$= p\left[\frac{\sum x_i^2}{\sigma_0^2} \leq \frac{\sigma^2{}_0}{\sigma_1^2} \chi_{\alpha,n}^2 / H_1\right]$$

$$= p\left[\chi^2 \leq \frac{\sigma^2{}_0}{\sigma_1^2} \chi_{\alpha,n}^2 / H_1\right]$$

Since under $H_1 \sim \frac{\sum x_i^2}{\sigma_1^2} \leq \chi^2_{(n)}$. Hence the power of the test is given by $F\left(\frac{\sigma^2{}_0}{\sigma_1^2}\chi^2_{\alpha,n}\right)$ where $F(.)$ is the distribution function of $\chi^2_{(n)}$.

Example 9: Examine whether test critical region exist for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ for parameter – of the distribution.

$$f(x,\theta) = \frac{1+\theta}{(x+\theta)^2}, 1 \leq x \leq \infty$$

Solution:

Let $x_1, x_2, ..., x_n$ be a random sample from exponential distribution then the likelihood function is given by

$$L_0 = \prod_{i=1}^{n} f(x_i, \theta_0) = \frac{(1+\theta_0)^2}{\prod_{i=1}^{n}(x_i + \theta_0)^2}$$

$$L_1 = \prod_{i=1}^{n} f(x_i, \theta_1) = \frac{(1+\theta_1)^2}{\prod_{i=1}^{n}(x_i + \theta_1)^2}$$

$$\frac{L_0}{L_1} = \frac{\dfrac{(1+\theta_0)^2}{\prod_{i=1}^{n}(x_i + \theta_0)^2}}{\dfrac{(1+\theta_1)^2}{\prod_{i=1}^{n}(x_i + \theta_1)^2}} \leq k$$

$$\left[\frac{(1+\theta_0)}{(1+\theta_1)}\right]^n \frac{\prod_{i=1}^{n}(x_i + \theta_1)^2}{\prod_{i=1}^{n}(x_i + \theta_0)^2} \leq k$$

Taking log on both sides,

$$n[\log(1+\theta_0) - \log(1+\theta_1)] + \sum_{i=1}^{n}\log(x_i + \theta_1)^2 - \sum_{i=1}^{n}\log(x_i + \theta_0)^2 \leq \log k$$

$$\sum 2\log\left(\frac{x_i + \theta_1}{x_i + \theta_0}\right) \leq \log k - n\log(1+\theta_0) + n\log(1+\theta_1)$$

Thus the test criterion is,

$$\sum_{i=1}^{n} 2\log\left(\frac{x_i+\theta_1}{x_i+\theta_0}\right)$$

This cannot be put up in the form of function of sample observations not depending upon the hypothesis. Hence, no BCR exits in this case.

Unbiased test:

A statistical test of simple null hypothesis against single alternative hypothesis is called unbiased if the power of the test is greater than or equal to single of the test.

Lemma for Unbiased Test:

The most powerful test for testing simple H_0 against simple H_1 is always unbiased. Let w be the best critical region of size for $H_0:\theta=\theta_0$ against $H_1:\theta=\theta_1$ from the population. Let $x_1,x_2,...,x_n$ be a random sample from $f(x,\theta)$ .let L_0 and L_1 be the likelihood function H_0 and H_1 respectively.

From Neymann Pearson Lemma (N-P Lemma) for the points inside w , $\frac{L_0}{L_1}\le k$ and for the point outside w , $\frac{L_0}{L_1}\ge k$ where k is a constant for the points inside w

$$L_0 \le k \;\; L_1$$

$$\int_w L_0 dx \;\; \le k \;\int_w L_1 dx$$

$$\alpha \le k \;\; (1-\beta) \qquad\qquad \text{------------}1$$

For the points outside w or the points inside $\overline{w}$

$$L_0 \ge k \;\; L_1$$

$$\int_{\overline{w}} L_0 dx \;\; \ge k \;\int_{\overline{w}} L_1 dx$$

$$1-\alpha \ge k \;\; \beta \qquad\qquad \text{-------------------}2$$

(1)* (2) gives

$$k \;\; (1-\beta)\;(1-\alpha) \ge k \;\; \beta\;\alpha$$

$$(1-\beta-\alpha+\alpha\beta) \ge \alpha\beta$$

$$1-\alpha-\beta+\alpha\beta-\alpha\beta \ge 0$$

$(1-\beta) \geq \alpha$

(ie) power of the test $\geq$ sign of the test.

Therefore the based on w is most powerful and unbiased.

Example 10: Obtain the most powerful test for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$ by taking a sample of size n from the normal population with known standard deviation σ_0.The pdf of normal distribution is given by.

$$f(x,\mu,\sigma) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma_0}\right)^2} \qquad -\infty \leq x,\mu \leq \infty, \sigma > 0$$

Solution

Let $x_1, x_2, ..., x_n$ be an random sample from normal population with the pdf let L_0 and L_1 be the likelihood function H_0 and H_1 respectively.

From NP lemma the best critical region is given by $w\ \{x_1, x_2 ..., x_n\}; \frac{L_0}{L_1} \leq k$ and the value of k can be found using.

$$p(x_1, x_2, ..., x_n)/ H_0 = \alpha$$

$$\frac{L_0}{L_1} \leq k$$

$$\frac{\prod_{i=1}^{n} f(x_i, \mu_0, \sigma_0)}{\prod_{i=1}^{n} f(x_i, \mu_1, \sigma_0)} \leq k$$

$$\frac{\prod_{i=1}^{n} \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_0}{\sigma_0}\right)^2}}{\prod_{i=1}^{n} \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_0}\right)^2}} \leq k$$

$$\frac{\left(\frac{1}{\sigma_0\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\left(\sum_{i=1}^{n}\left(\frac{x_i-\mu_0}{\sigma_0}\right)^2\right)}}{\left(\frac{1}{\sigma_0\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\left(\sum_{i=1}^{n}\left(\frac{x_i-\mu_1}{\sigma_0}\right)^2\right)}} \leq k$$

$$\frac{e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\left(\frac{x_i-\mu_0}{\sigma_0}\right)^2}}{e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\left(\frac{x_i-\mu_1}{\sigma_0}\right)^2}}\le k$$

$$e^{\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{n}\left(\frac{x_i-\mu_0}{\sigma_0}\right)^2-\sum_{i=1}^{n}\left(\frac{x_i-\mu_1}{\sigma_0}\right)^2\right]}\le k$$

Taking log on both sides,

$$-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{n}(x_i-\mu_0)^2-\sum_{i=1}^{n}(x_i-\mu_1)^2\right]\le \log k$$

$$\sum_{i=1}^{n}(x_i-\mu_1)^2-\sum_{i=1}^{n}(x_i-\mu_0)^2\le 2\sigma_0^2\log k$$

$$\sum_{i=1}^{n}(x_i-\mu_1)^2-\sum_{i=1}^{n}(x_i-\mu_0)^2\le c$$

$$\sum_{i=1}^{n}\left[x_i^2+\mu_1^2-2x_i\mu_i-x_i^2-\mu_0^2+2x_i\mu_0\right]\le c$$

$$\sum_{i=1}^{n}\left[\mu_1^2-\mu_0^2-2x_i(\mu_1-\mu_0)\right]\le c$$

$$n(\mu_1^2-\mu_0^2)-2(\mu_1-\mu_0)\sum_{i=1}^{n}x_i\le c$$

$$-(\mu_1-\mu_0)\sum_{i=1}^{n}x_i\le c\left(\frac{n}{2}\right)(\mu_1^2-\mu_0^2)$$

$$(\mu_1-\mu_0)\sum_{i=1}^{n}x_i\le c_1$$

$w\,\{x_1,x_2...,x_n\};(\mu_1-\mu_0)\sum_{i=1}^{n}x_i\le c_1$ is the best critical region and can be obtained using

$$p(x_1,x_2,...,x_n)/H_0=\alpha$$

Case I: $\mu_0<\mu_1$

$\mu_0-\mu_1$ is a negative quantity. Dividing best critical region of equation (1) $\mu_0-\mu_1$ we get

$$\sum_{i=1}^{n}x_i\ge\frac{c_1}{\mu_0-\mu_1}=c_2$$

$$\bar{x} \geq 3 \, where \; c_3 = \frac{c_2}{n}$$

Case II: $\mu_0 > \mu_1$

$\mu_0 - \mu_1$ is a negative quantity. Dividing best critical region of equation (1) $\mu_0 - \mu_1$ we get

$$\sum_{i=1}^{n} x_i \leq \frac{c_1}{\mu_0 - \mu_1} = c_2$$

$$\bar{x} = 3 \, where \; c_3 = \frac{c_2}{n}$$

Example 11: Obtain the most powerful test size α for testing $H_0 : \sigma = \sigma_0$ versus $H_1 : \sigma = \sigma_1$ in $N(0, \sigma_0^2)$. Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from normal population with mean 0 and variance σ^2 .

Solution:

The likelihood function for $N(0, \sigma_0^2)$ is

$$\prod_{i=1}^{n} f(x_i, 0, \sigma_0^2) = \left(\frac{1}{\sigma_0^{\;2} \sqrt{2\pi}} \right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - 0}{\sigma_0}\right)^2}$$

$$= \left(\frac{1}{\sigma_0^{\;2} \, 2\pi} \right)^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{x_i^2}{\sigma_0^2}}$$

The likelihood function for $N(0, \sigma_1^2)$ is

$$\prod_{i=1}^{n} f(x_i, 0, \sigma_1^2) = \left(\frac{1}{\sigma_1^{\;2} \sqrt{2\pi}} \right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - 0}{\sigma_1}\right)^2}$$

The best critical region is given by

$$\frac{L_0}{L_1} \leq k$$

$$\frac{\left(\frac{1}{\sigma_0^{\;2} \, 2\pi} \right)^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{x_i^2}{\sigma_0^2}}}{\left(\frac{1}{\sigma_1^{\;2} \, 2\pi} \right)^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{x_i^2}{\sigma_1^2}}} \leq k$$

$$\left(\frac{\sigma_1^2}{\sigma_0^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\left[\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right]} \le k$$

$$\left(\frac{\sigma_1}{\sigma_0}\right)^{n} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\left[\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right]} \le k$$

Taking log on both sides

$$n\log\left(\frac{\sigma_1}{\sigma_0}\right) \; -\frac{1}{2}\sum_{i=1}^{n} x_i^2\left[\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right] \le \log k$$

$$n\left[\log \sigma_1 - \log \sigma_0\right] - \frac{1}{2}\sum_{i=1}^{n} x_i^2\left[\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right] \le \log k$$

$$-\sum_{i=1}^{n} x_i^2\left[\sigma_1^2-\sigma_0^2\right] \le \log k - n\left[\log \sigma_1 - \log \sigma_0\right]2\sigma_1^2 * \sigma_0^2$$

$$\sum_{i=1}^{n} x_i^2\left[\sigma_0^2-\sigma_1^2\right] \le \left[\log k - n\log \sigma_1 + \log \sigma_0\right]2\sigma_1^2 * \sigma_0^2$$

$$\sum_{i=1}^{n} x_i^2\left[\sigma_0-\sigma_1\right] \le \frac{\left[\log k - n\log \sigma_1 + \log \sigma_0\right]2\sigma_1^2 * \sigma_0^2}{\sigma_0+\sigma_1}$$

$$\sum_{i=1}^{n} x_i^2\left[\sigma_0-\sigma_1\right] \le c$$

Case I: $\sigma_0 < \sigma_1$

Critical region is $\sum_{i=1}^{n} x_i^2\left[\sigma_0-\sigma_1\right] \le c$

$$\sum_{i=1}^{n} x_i^2 \ge \frac{c}{\sigma_0-\sigma_1} \qquad \Rightarrow \sum_{i=1}^{n} x_i^2 \ge c_1$$

Case II: $\sigma_0 > \sigma_1$

$$\sum_{i=1}^{n} x_i^2 \le \frac{c}{\sigma_0-\sigma_1} \qquad \Rightarrow \sum_{i=1}^{n} x_i^2 \le c_1$$

Example 12: Obtain the most powerful test size α for testing $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda = \lambda_1$ from poison population for the parameter λ .

Solution:

The probability mass function of poison population distribution is given by

$$p[x;\lambda]=\frac{e^{-\lambda}\lambda^{x}}{x!}$$

Let $x_1, x_2,...,x_n$ be a random sample from poison distribution then the likelihood function is given by

$$\prod_{i=1}^{n}\frac{e^{\lambda_0}\lambda_1^{x_i}}{x_i!}$$

$$L_0 = \frac{e^{-n\lambda_0}\lambda_0^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}$$

$$L_1 = \frac{e^{-n\lambda_1}\lambda_1^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}$$

$$\frac{L_0}{L_1}\leq k$$

$$\frac{\dfrac{e^{-n\lambda_0}\lambda_0^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}}{\dfrac{e^{-n\lambda_1}\lambda_1^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}}\leq k$$

$$e^{-n(\lambda_0-\lambda_1)}\left(\frac{\lambda_0}{\lambda_1}\right)^{\sum_{i=1}^{n}x_i}\leq k$$

Taking log on both sides

$$-n(\lambda_0-\lambda_1)\log e+\sum_{i=1}^{n}x_i\log\left(\frac{\lambda_0}{\lambda_1}\right)\leq \log k$$

$$-n(\lambda_0-\lambda_1)+\sum_{i=1}^{n}x_i(\log\lambda_0-\log\lambda_1)\leq c$$

$$\sum_{i=1}^{n}x_i(\log\lambda_0-\log\lambda_1)\leq c+n(\lambda_0-\lambda_1)$$

Case I: $\lambda_0 > \lambda_1$

$$\log \lambda_0 - \log \lambda_1 \text{ is a positive quantity}$$

The inequality remains same the best critical region is given by

$$\sum_{i=1}^{n} x_i \leq \frac{c + n(\lambda_0 - \lambda_1)}{\log \lambda_0 - \log \lambda_1}$$

Case II: $\lambda_0 < \lambda_1$

Since $\log \lambda_0 - \log \lambda_1$ is a negative quantity.

The inequality best critical region is given by

$$\sum_{i=1}^{n} x_i \geq \frac{c + n(\lambda_0 - \lambda_1)}{\log \lambda_0 - \log \lambda_1}$$

Example 13: obtain the best critical region of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ in the exponential population. The probability density function of exponential distribution is given by.

$$f(x, \theta) = \theta e^{-\theta x}$$

Solution:

Let $x_1, x_2, ..., x_n$ be a random sample from exponential distribution then the likelihood function is given by

$$\prod_{i=1}^{n} f(x_i, \theta_0) = \prod_{i=1}^{n} \theta_0 e^{-\theta_0 x_i}$$

$$\frac{L_0}{L_1} = \frac{\theta_0{}^n e^{-\theta_0 \sum_{i=1}^{n} x_i}}{\theta_1{}^n e^{-\theta_1 \sum_{i=1}^{n} x_i}} \leq k$$

$$\left(\frac{\theta_0}{\theta_1}\right)^n e^{-\sum_{i=1}^{n} x_i (\theta_0 - \theta_1)} \leq k$$

Taking log on both sides,

$$n \log\left(\frac{\theta_0}{\theta_1}\right) - \sum_{i=1}^{n} x_i (\theta_0 - \theta_1) \leq \log k$$

$$n[\log \theta_0 - \log \theta_1] - \sum_{i=1}^{n} x_i (\theta_0 - \theta_1) \leq \log k$$

$$-\sum_{i=1}^{n} x_i (\theta_0 - \theta_1) \leq \log k - n[\log \theta_0 - \log \theta_1]$$

Case I: $\theta_0 > \theta_1$

$$\Rightarrow \theta_0 - \theta_1 > 0$$

$\log \theta_0 - \log \theta_1$ is a positive quantity

$$\sum_{i=1}^{n} x_i \geq \frac{\log k - n[\log \theta_0 - \log \theta_1]}{\theta_0 - \theta_1}$$

Case II: $\theta_0 < \theta_1$

$\log \theta_0 - \log \theta_1$ is a positive quantity

$$\sum_{i=1}^{n} x_i \leq \frac{\log k + n[\log \theta_0 - \log \theta_1]}{\theta_0 - \theta_1}$$

Example 14: Obtain the most powerful test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ for the pdf $f(x) = \theta x^{\theta - 1}$ $0 < x < 1$ $, \theta \geq 1$

Solution:

Let $x_1, x_2, ..., x_n$ be a random sample.

$$\frac{L_0}{L_1} \leq k$$

$$\frac{\prod_{i=1}^{n} \theta_0 x_i^{\theta_0 - 1}}{\prod_{i=1}^{n} \theta_1 x_i^{\theta_1 - 1}} \leq k$$

$$\frac{\theta_0 \prod_{i=1}^{n} x_i^{\theta_0 - 1}}{\theta_1 \prod_{i=1}^{n} x_i^{\theta_1 - 1}} \leq k$$

$$\left(\frac{\theta_0}{\theta_1}\right)^n \prod_{i=1}^{n} x_i^{(\theta_0 - 1 - \theta_1 + 1)} \leq k$$

$$\left(\frac{\theta_0}{\theta_1}\right)^n \prod_{i=1}^{n} x_i^{(\theta_0 - \theta_1)} \leq k$$

Taking log on both sides,

$$n[\log \theta_0 - \log \theta_1] + \log \prod_{i=1}^{n} x_i^{(\theta_0 - \theta_1)} \leq \log k$$

$$n[\log\theta_0 - \log\theta_1] + (\theta_0 - \theta_1)\sum_{i=1}^{n}\log x_i \le \log k$$

$$\sum_{i=1}^{n}\log x_i \le \frac{\log k - n[\log\theta_0 - \log\theta_1]}{(\theta_0 - \theta_1)}$$

Case I: $\theta_0 > \theta_1$

$$\Rightarrow \theta_0 - \theta_1 > 0$$

$\log\theta_0 - \log\theta_1$ is a positive quantity

$$\sum_{i=1}^{n} x_i \le \frac{\log k - n[\log\theta_0 - \log\theta_1]}{\theta_0 - \theta_1}$$

Case II: $\theta_0 > \theta_1$

$\log\theta_0 - \log\theta_1$ is a positive quantity

$$\sum_{i=1}^{n} x_i \ge \frac{\log k - n[\log\theta_0 - \log\theta_1]}{\theta_0 - \theta_1}$$

LIKELIHOOD RATIO TEST:

Likelihood ratio test is useful for testing simple or composite hypothesis. If $f(x,\theta)$ is the density function of a population and $L(\theta)$ is a likelihood function of sample observations $x_{1,}$ x_2, x_3,…, x_n then the likelihood ratio λ is defined as

$$\lambda = \frac{Maximum\ of\ Likelihood\ function\ \ L(\theta)|H_0}{Maximum\ of\ \ L(\theta)}$$

If the parameter θ is replaced by its maximum likelihood estimator $\hat{\theta}$, then we get $L(\hat{\theta})$. ie., $H_0 : \theta = \theta_0$, then we get $L(\hat{\theta})$. (ie) Max $L(\theta) = L(\hat{\theta})$

$$\lambda = \frac{L(\theta_0)}{L(\hat{\theta})}$$

Any test for testing H_0 against H_1 is called likelihood ratio test. If it is based on likelihood ratio λ and the critical region $0 \le \lambda \le \lambda_0$ such that $\int_0^{\lambda} g(\lambda|H_0)d\lambda = \alpha$

Properties of Likelihood Ratio Test

1. Likelihood ratio test leads to uniformly most powerful test if it exists.
2. When the sample size n is large $-2\log_e \lambda \sim \chi^2$ distribution with respective degrees of freedom
3. Under certain conditions likelihood ratio tests are consistent.
4. If the distribution $f(x,\theta)$ has a monotone likelihood ratio in D(x) then there exists UMP test for testing $H_0 : \theta \le \theta_0$ or $H_0 : \theta \ge \theta_0$ against $H_1 : \theta < \theta_0$

Example 15: Obtain UMPT(LRT) for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \ne \mu_0$ for a normal population with parameter μ and σ^2.

Solution:

Let x

Let $x_1,x_2,\ldots,x_n$ be a random sample from N(μ , σ^2) where x, $\mu \in \Re$, $\sigma > 0$

The joint pdf of $x_1,x_2,\ldots,x_n$ is

$$\prod_{i=1}^{n} f\left(x_i : \mu,\sigma^2\right) = \prod_{i=1}^{n}\left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right], x_i, \mu \in \Re, \sigma > 0$$

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

$$L\left(\mu_0,\sigma^2\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum\left(\frac{x_i-\mu_0}{\sigma}\right)^2}$$

MLE of μ and σ^2 are $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum\left(x_i - \bar{x}\right)^2 = s^2$

The maximum of likelihood function is given by

$$L\left(\hat{\mu},\hat{\sigma}^2\right) = \left(\frac{1}{s\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum\left(\frac{x_i-\bar{x}}{s}\right)^2} = \left(2\pi s^2\right)^{-n/2} e^{-n/2} \text{----(1)}$$

Maximum Likelihood estimator of for σ^2 when $H_0 : \mu = \mu_0$ is true given by

$$\sigma^2 = \frac{1}{n}\sum\left(x_i - \mu_0\right)^2 = \frac{1}{n}\sum\left(x_i - \bar{x} + \bar{x} - \mu_0\right)^2 = s^2 + \left(\bar{x} - \mu_o\right)^2$$

Therefore, $\hat{\sigma}^2 = s_0^2$

Maximum likelihood function under H_o is

$$L\left(\hat{\sigma}^2 \middle| H_0\right)=\left(\frac{1}{s_0\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum\left(\frac{x_i-\bar{x}}{s_o}\right)^2} = \left(2\pi s_o^2\right)^{-n/2} e^{-n/2}\text{----(2)}$$

The UMP critical region of size α is given by $0 \le \lambda \le \lambda_0$

$$\Rightarrow \frac{Maximum \;\; Likelihood \;\; function|H_o}{Maximum \;\; Likelihood \;\; function} \le \lambda_0$$

Using (1) and (2),

$$\Rightarrow \frac{\left(2\pi s_o^2\right)^{-n/2} e^{-n/2}}{\left(2\pi s^2\right)^{-n/2} e^{-n/2}} \le \lambda_0$$

$$\Rightarrow \left(\frac{s_0^2}{s^2}\right)^{-n/2} \le \lambda_0 \Rightarrow \left(\frac{s^2+\left(\bar{x}-\mu_o\right)^2}{s^2}\right)^{-n/2} \le \lambda_0 \Rightarrow \left(\frac{1}{1+\frac{\left(\bar{x}-\mu_o\right)^2}{s^2}}\right)^{n/2} \le \lambda_0$$

where λ_o is fixed such that size of CR is

$$t=\frac{\left(\bar{x}-\mu_0\right)}{s/\sqrt{n-1}} \sim t_{n-1}$$

$$\lambda=\left[\frac{1}{1+\frac{t^2}{n-1}}\right]^{n/2} \le \lambda_0$$

Therefore, t-distribution can be used to find the value for given α and degrees of freedom (n-1).

Therefore, UMPT of size α for testing mean of the normal distribution when σ^2 is unknown is based on t-distribution.

The UMP CR of Size α is given by

$$\left[\frac{1}{1+\frac{t^2}{n-1}}\right]^{n/2} \le \lambda_0 \Rightarrow \left[\frac{1}{1+\frac{t^2}{n-1}}\right] \le \lambda_0^{2/n}$$

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

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

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

Test for the Mean of a Normal Distribution

Let X$_1$,X$_2$,...,X$_n$ form a random sample from a normal distribution whose mean μ and variance σ^2 are both unknown. Consider the problem of testing the composite null hypothesis $\mathrm{H}_0 : \mu = \mu_0$ against $\mathrm{H}_1 : \mu = \mu_1$.

The joint probability density function of X$_1$,X$_2$,...,X$_n$ under H$_0$, where σ^2 is regarded as the parameter, is

$$f(X \mid \mu_0, \sigma^2) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{-\frac{\sum_i (x_i - \mu_0)^2}{2\sigma^2}}$$

This shows that the statistic, $V = \sum_i (X_i - \mu_0)^2$ is sufficient for σ^2 and also complete sufficient statistic.

Consider now a particular simple hypothesis,

$$H_0 : \mu = \mu_0, \sigma^2 = \sigma_0^2 \text{ and}$$

$H_1 : \mu = \mu_1, \sigma^2 = \sigma_1^2$

The most powerful similar region of size α for testing H$_0$ against H$_1$ is

$W_0 = \{x \mid f(x \mid \mu_1, \sigma_1^2) > k(v) f(x \mid \mu_0, \sigma_0^2)\}$

where k(v) is such that the conditional size of W$_0$ given V=v, is α.

Now, if we take logarithms on both sides, we see that

$f(x \mid \mu_1, \sigma_1^2) > k(v) f(x \mid \mu_0, \sigma_0^2)$ iff

$(\mu_1 - \mu_0)(\bar{x} - \mu_0) > k_1(v)$, ---(1)

say, where k$_1$(v) is related to k(v).

Case 1: $\mu_1 > \mu_0$

Hence the condition (1) is equivalent to

$(\bar{x} - \mu_0) > k_2(v)$ or to

$\sqrt{n}(\bar{x} - \mu_0)/\sqrt{v} > k_3(v),$

As such, we may write

$W_0 = \{x \mid \sqrt{n}(\bar{x}-\mu_0)/\sqrt{v} > k_3(v)\}$

Where, again k3(v) is to be so determined that $P_{\theta_0}(W_0 \mid v) = \alpha$

However, $\sqrt{n}(\bar{x}-\mu_0)/\sqrt{V}$ and V are independently distributed, so that the con0ditional distribution $\sqrt{n}(\bar{x}-\mu_0)/\sqrt{V}$, given V=v, is the same as the marginal distribution of $\sqrt{n}(\bar{x}-\mu_0)/\sqrt{v}$. Such, k3(v) will be independent of v. Writing k3 for this constant, we see that it is to be so determinant that

$P_{\theta_0}\left[\sqrt{n}(\bar{x}-\mu_0)/\sqrt{V} > k_3\right] = \alpha$

We also note that,

$$\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sqrt{V}} = \frac{\sqrt{n}(\bar{x}-\mu_0)}{\left\{n(\bar{X}-\mu_0)^2 + \sum_i (X_i-\bar{X})^2\right\}^{1/2}} = \frac{1}{\sqrt{t^2+n-1}}$$

where $t = \sqrt{n}(\bar{X}-\mu_0)/S$ is Student's t statistic n-1degrees of freedom.

Since $\sqrt{n}(\bar{X}-\mu_0)/\sqrt{V} > k_3$ iff $t > k_4(say)$

We may also write, $W_0 = \{x \mid \sqrt{n}(\bar{x}-\mu_0)/\sqrt{v} > k_3\} = \{x \mid t > k_3\}$

where k4 is such that, $P_{\theta_0}[t > k_4] = \alpha$

This shows that k4 is the upper $\alpha-\text{point}$ of the t distribution with n-1 degrees of freedom. Denoting this by $t_{\alpha,n-1}$, then

$W_0 = \{x \mid \sqrt{n}(\bar{x}-\mu_0)/\sqrt{s} > t_{\alpha,n-1}\}$

Since this is independent of σ_0^2, it is the most powerful similar region of size α for testing H$_0$ against H$_1$.

Case 2: $\mu_1 < \mu_0$

In this case, condition (1) reduces to

$(\bar{x}-\mu_0) < k_2'(v)$

Proceeding as before, we shall find the most powerful similar region of size α for testing H$_0$ against H$_1$ is

$W_0' = \{x \mid \sqrt{n}(\bar{x}-\mu_0)/s < -t_{\alpha,n-1}\}$

Note: Since W$_0$ is independent of μ_1, i.e., is the same for all $\mu_1 > \mu_0$, it is, in fact, the UMP similar region of Size α for testing $H_0 : \mu = \mu_0$ against the more general composite alternative $H_0 : \mu > \mu_0$. Similarly, $W_0^{'}$ is the uniformly most powerful similar region of size α for testing $H_0 : \mu = \mu_0$ against the alternative $H_0 : \mu < \mu_0$.

Test for Variance of a Normal Distribution

Let X$_1$,X$_2$,...,X$_n$ form a random sample from a normal distribution whose mean μ and variance σ^2 are both unknown. Consider the problem of testing the composite null hypothesis

$\mathrm{H}_0 : \sigma = \sigma_0$ against

$\mathrm{H}_1 : \sigma = \sigma_1$.

The joint probability density function of X$_1$,X$_2$,...,X$_n$ under H$_0$, where μ is regarded as the parameter, is

$$f(X \mid \mu, \sigma_0^2) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{-\frac{\sum_i (x_i - \mu)^2}{2\sigma_0^2}}$$

This shows that the statistic, $\overline{X} = \sum_i X_i / n$ is sufficient for μ and also complete sufficient statistic under H$_0$.

Consider now a particular simple hypothesis,

$$H_0 : \mu = \mu_0, \sigma^2 = \sigma_0^2 \text{ and}$$

$H_1 : \mu = \mu_1, \sigma^2 = \sigma_1^2$

The most powerful similar region of size α for testing H$_0$ against H$_1$ is

$W_0 = \{x \mid f(x \mid \mu_1, \sigma_1^2) > k(\bar{x}) f(x \mid \mu_0, \sigma_0^2)\}$

where $k(\bar{x})$ is such that the conditional size of W$_0$ given $\overline{X} = \bar{x}$, is α.

Now, the condition, $f(x \mid \mu_1, \sigma_1^2) > k(\bar{x}) f(x \mid \mu_0, \sigma_0^2)$

Reduces, if we take logarithms on both sides, to

$(\sigma_1^2 - \sigma_0^2) \sum_i (x_i - \bar{x})^2 > k_1(\bar{x})$ (say) ---(1)

Case I: $\sigma_1^2 > \sigma_0^2$

Here condition (1) is equivalent to $\sum_i (x_i - \bar{x})^2 > k_2(\bar{x})$ (say) or $\sum_i \frac{(x_i - \bar{x})^2}{\sigma_0^2} > k_3(\bar{x})$(say)

we may, therefore, write

$$W_0 = \left\{ x \mid \sum_i \frac{(x_i - \bar{x})^2}{\sigma_0^2} > k_3(\bar{x}) \right\}$$

Where, $k_3(\bar{x})$ is to be so determined that

$$P_{\theta_0}(W_0|\bar{x}) = \alpha$$

Since $\sum_i \frac{(X_i - \bar{X})^2}{\sigma_0^2}$ and $\bar{X}$ are independently distributed, the conditional distribution of $\sum_i \frac{(x_i - \bar{x})^2}{\sigma_0^2}$ given $\bar{X} = \bar{x}$ is the same as its marginal distribution, implying that $k_3(\bar{x})$ is independent of $\bar{x}$.

Writing k_3 for this constant, we note that it is to be so chosen that

$$P_{\theta_0}\left(\sum_i \frac{(X_i - \bar{X})^2}{\sigma_0^2} > k_3 \right) = \alpha$$

Since $\sum_i \frac{(X_i - \bar{X})^2}{\sigma_0^2}$ has, under H$_0$, the χ^2 distribution with n-1 degrees of freedom, k$_3$ must be upper $\alpha -$point of the χ^2 distribution with n-1 degrees of freedom. Denoting this by $\chi^2_{\alpha,n-1}$, then

$$W_0 = \left\{ x \mid \frac{(X_i - \bar{X})^2}{\sigma_0^2} > \chi^2_{\alpha,n-1} \right\}$$

Since this is independent of μ_0 and μ_1, it is the most powerful similar region of size α for testing H$_0$ against H$_1$.

Case II: $\sigma_1^2 < \sigma_0^2$

In this case, condition (1) reduces to

$$\sum_i \frac{(X_i - \bar{X})^2}{\sigma_0^2} < k_2^{'}(\bar{x}).$$

Proceeding as before, we shall find the most powerful similar region of size α for testing H$_0$ against H$_1$ is

$$W_0^{'} = \left\{ x \mid \frac{\left(X_i - \overline{X}\right)^2}{\sigma_0^2} < \chi^2_{1-\alpha,n-1} \right\}$$

Where $\chi^2_{1-\alpha,n-1}$ is the lower $\alpha - po$int of the χ^2 distribution with n-1 degrees of freedom.

Note: Since W$_0$ is the same for all $\sigma_1^2 > \sigma_0^2$, it is, in fact, the uniformly most powerful similar region of size α for testing H$_0$ against the one-sided alternative $H_1 : \sigma^2 > \sigma_0^2$. Similarly, $W_0^{'}$ is the uniformly most powerful similar region of size α for testing H$_0$ against the one-sided alternative $H_1 : \sigma^2 < \sigma_0^2$.

UNIT-IV

In the previous chapter, we have discussed the statistical hypothesis. In this chapter, we shall discuss the large sample tests, exact sample tests and chi-square tests.

LARGE SAMPLE TEST:

Any statistical test based on the assumption that the sample size n is large (n$\rightarrow \infty$) is called asymptotic test. We know that as n$\rightarrow \infty$any statistic irrespective of the parent population from which sample is drawn follows Normal Distribution (Central limit theorem).

Hence any statistic follows Normal Distribution as n$\rightarrow \infty$the test based on such a statistic is called asymptotic test. Any statistical test based on exact distribution of a statistical under consideration is called exact test. Here, there is no assumption on the sample size most of the statistical test uses t-distribution, x^2 distribution and F-distribution which are exact distribution of statistic. Hence test based on t, F, x^2 distributions are called exact test. Sometimes the statistic may also follow Normal Distribution and in such cases, it is also an exact test.

Steps involved in statistical test of significance:

A statistical test of significance is a statistical test of hypothesis using the following procedure.

1. Formulation of hypothesis:

The hypothesis to be test is taken as null hypothesisH_0. Normally when one parameter is involved the hypothesis is "there is no significant difference between the hypothetical value of the parameter and corresponding statistical value from the sample". When two parameters are involved, the null hypothesis is "there is no significant difference between statistic obtained from two sample". The alternative hypothesis is normally two sided and just opposite of null hypothesis.

2. Chossing the level of significance:

α = level of significance

= P [Type I error]

= size of critical region

α value is fixed at low level usually it is fixed as 5% or 1%.

3. Selecting statistic & finding its distribution:

Let t be a statistic such that E(t)= ϕ where ϕ is the parameter of the distribution. We must find standard error of t which is the standard deviation of the sampling distribution of the statistic.

test statistic = $\frac{t - E(t)}{SE(t)}$

Find the distribution of test statistic which may be normal, t, x^2 or F distribution.

4. Finding the critical value:

Using the sampling distribution of test statistic critical value or table value can be obtained from the corresponding statistical tables. These values are used to describe the critical region. For eg. If the sampling distribution is normal, normal table can be use to find critical value using α. And if α = 0.05, H_1 is two sided, then the critical value is 1.96. If α = 0.05, H_1 is one sided , then the critical value is 1.965. If α =0.01 and H_1 is two sided, the critical value is 2.58 and if H_1 is one sided, the critical value is 2.33.

5. Critical region & inference:

Critical region is { |test statistic| $\geq$ critical values }

A.R = { |test statistic| < critical values}.

If the value of test statistic $\geq$ critical value H_1 is rejected. If the value of test statistic < critical value then there is no reason to reject H_0 at level α. Accordingly, inferences can be drawn.

TEST FOR SINGLE MEAN:

Give the test procedure for testing the significance of mean of the population when the sample is large.

Null hypothesis: There is no significant difference between sample mean and population (i.e) $H_0 : \mu = \mu_0$.

Alternate hypothesis: There is significant difference between sample mean and population mean (i.e) $H_1 : \mu \neq \mu_0\ (or)\ H_1 : \mu < \mu_0\ (or)\ H_1 : \mu > \mu_0$

Level of Significance: Let == be the level of significance, == 0.05 or 0.01 or any given specific values in the problem.

Test statistic & its sampling distribution:

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \quad \text{where } \sigma \text{ is known}$$

$$Z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \quad \text{where } \sigma \text{ is unknown}$$

Where n is sample size

$\bar{x}$ is sample mean

σ is population SD

s is sample SD

Finding critical value:

From the normal table, we find critical value based on α & H_1 is one-sided then critical value is Z_α, if the H_1 is two-sided then critical value is $Z_{\alpha/2}$.

Inference:

If $|Z_{cal}| > Z_\alpha\ (or\ Z_{\alpha/2})$ then the null hypothesis is rejected.

If $|Z_{cal}| \leq Z_\alpha\ (or\ Z_{\alpha/2})$ then there is no reason to reject it.

Problem 1: A sample of 900 members has a mean 3.4 cm and SD 2.61 cm is a sample from a large population of mean 3.25 cm & SD 2.61 cm.

Solution:

n=900 $\bar{x}$ =3.4cm μ_0= 3.25 σ= 2.61

Null hypothesis:

The sample has drawn from the population with mean 3.25cm & SD=2.61 cm

Alternate hypothesis:

The sample is not drawn from the population with mean 3.25cm & SD=2.61 cm

$$Z_{cal} = \frac{\bar{x}-\mu_0}{\sigma / \sqrt{n}}$$

$$Z = \frac{3.4-3.25}{2.61/\sqrt{900}} = \frac{0.15}{(2.61)/30} = \frac{(0.15)30}{2.61}$$

Z= 1.7241

Critical value:

Let α =0.05 , $\alpha/2$ = 0.025 $Z_{0.025}$ =1.96

Inference:

Since Z_{cal} = 1.7241 < Z_{cal} = 1.96

$\Rightarrow$ There is no reason to reject null hypothesis.

We conclude that sample has been drawn from population with mean 3.25 & SD= 2.61 cm.

Problem 2: An insurance agent has claimed that the average age of policy holders who insure through him is less than the average for all agent which is 30.5 years. A random sample of 100 policy holders who had insured through him gave the following distribution.

Age	15-20	20-25	25-30	35-40	30-35
No.of.persons	12	22	20	16	30

Calculate the AM and SD of this distribution &use these values to test thi claim at 5% level of significance.

Solution:

CI	x_i	f_i	$x_i f_i$	$x_i^2 f_i$
15-20	17.5	12	210	3675
20-25	22.5	22	495	11137.5
25-30	27.5	20	550	15125
30-35	32.5	30	975	31687.5
35-40	37.5	16	600	22500

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{2830}{100} = 28.3$$

$$\sigma^2 = \frac{1}{N}\Sigma x_i^2 f_i - (\bar{x})^2$$

$= 841.25-800.89$

$= 40.36$

$\sigma = s = 6.353$

Null hypothesis:

The average age of policy holders who insured through him is same as the average age for all agents which is 30.5 years.

Alternate hypothesis:

The average age of policy holders who are insured through him is less than the average age for all agents which is 30.5 years

Test statistic:

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$= \frac{28.3 - 30.5_0}{6.353/\sqrt{100}}$$

$$= \frac{-2.2}{0.6353}$$

$= -3.4629$

$Z_{\alpha=1.65} = Z_{tab}$

$|Z_{cal}| = |3.4629| > |1.65| = Z_{tab}$

Therefore we reject the null hypothesis and accept that him claim is right.

Since $|Z_{cal}| = 3.4629 > Z_{\alpha} = 1.65$

H_0 is rejected. Thus we conclude that the insurance agents claim is true. (i.e) Average age of policy holders who insured through him is less than the average age for all agents which is 30.5 years.

TEST OF SIGNIFICANCE OF STANDARD DEVIATION OR VARIANCE:

Null hypothesis:

$$H_0 : \sigma = \sigma_0 \ (or) \ \sigma^2 = \sigma_0^2$$

There is no significant difference between sample variance and population variance.

Alternate hypothesis:

$H_1 : \sigma \neq \sigma_0 \ (or) \ \sigma^2 \neq \sigma_0^2 \ or$

$H_1 : \sigma > \sigma_0 \ (or) \ \sigma^2 > \sigma_0^2 \ or$

$H_1 : \sigma < \sigma_0 \ (or) \ \sigma^2 < \sigma_0^2$

Level of significance:

α = P[Type I error]= 0.05/001

Test statistic:

$$Z = \frac{s - E(s)}{s.E(s)}$$

Under H_0, $Z = \dfrac{s - \sigma_0}{\sigma_0 / \sqrt{2n}} \approx N(0,1)$

Where n is the sample size

$s = \sqrt{\dfrac{1}{n}\sum(x_i - \bar{x})^2}$ sample SD

$\sigma_0 =$ population SD

Critical values:

$H_1 : \sigma \neq \sigma_0$

From normal distribution table, we can find

$Z_{\alpha/2}$ such that

$P[Z_{cal} > Z_{\alpha/2}] = \alpha/2$

$H_1 : \sigma > \sigma_0$

From normal distribution table, we can find Z_α Such that

$P[Z_{cal} > Z_\alpha] = \alpha$

$H_1 : \sigma < \sigma_0$

From normal distribution table, we can find - Z_α Such that

$P[Z_{cal} < Z_\alpha] = \alpha$

Inference:

When $H_1 : \sigma_1 \neq \sigma_2$ reject H_0 if $|Z_{cal}| > Z_{\alpha/2}$ otherwise there is no reason to reject H_0.

When $H_1 : \sigma_1 < \sigma_2 \, (or) \, H_1 : \sigma_1 > \sigma_2$ reject H_0 if $|Z_{cal}| < Z_\alpha \, (or) |Z_{cal}| > Z_\alpha$ otherwise there is no reason to reject H_0.

Problem 3: A large organisation produces electrical light bulbs in each of its two factories. It is suspected that the efficiency of the factories are not same. So a test carried out by ascertaining variability of life of bulbs produced in each factory. The results are as follows:

No.of bulbs in the sample	Factory A	Factory B
	100	200
Average life	1100 hrs	900 hrs
SD	240 hrs	220 hrs

From the above information determine whether the difference between variability of life from bulbs from each sample is significant. Test at 5% level of significance.

Solution:

Null hypothesis:

There is no significant difference between the variability of life of bulbs from factory A and factory B.

Alternative hypothesis:

There is no significant difference between the variability of life of bulbs in factory A and factory B.

Level of significance:

α = P [Type I error]= 0.01

Test statistic:

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s\sigma_1^2}{2n_1} + \frac{s\sigma_2^2}{2n_2}}}$$

$$= \frac{(240)^2 - (220)^2}{\sqrt{\frac{(240)^2}{200} + \frac{(220)^2}{400}}}$$

$$= \frac{20}{\sqrt{288 + 121}} = \frac{20}{20.2237} = 0.9889$$

$Z_{cal} = 0.9889$

$$Z_{tab} = 0.01 \Rightarrow Z_{\alpha/2} = 2.58$$

$$Z_{cal} = 0.9889 < Z_{tab} = 2.58$$

Inference: There is no reason reject H_0. There is no significant difference between variability of life of bulbs of factory A and factory B.

TEST FOR SIGNIFICANCE OF SAMPLE PROPORTION:

Let $x_{1,}x_{2,}......,x_n$ be a sample observation of size n with the proportion p, q=1-p. We have to test there is any significant difference between the sample proportion (p) and population proportion (p) where n is assumed to be large.

Null hypothesis:

There is no significant difference between the sample proportion and population proportion H_0: $p=p_0$

Alternative hypothesis:

There is no significancant difference between the sample proportion and population proportion.

$H_1 : p \neq p_0$

$H_1 : p < p_0$

$H_1 : p > p_0$

Level of significance:

α = p [Type of I error]=0.01/0.05 or any other specified values.

Test statistic:

$$H_0, Z = \frac{p - E(p)}{s\sqrt{v(p)}} \approx N(0,1)$$

Under

$$= \frac{p - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

Critical value:

When $H_1 : p \neq p_0$ from normal table we can find $Z_{\alpha/2}$ using $p[Z > Z_{\alpha/2}] = \alpha/2$

When $H_1 : p < p_0 \, or \, p > p_0$ from normal table we can find $Z_{\alpha/2}$ using $p[Z \leq Z_\alpha] = \alpha \, (or) \, p[Z \geq Z_\alpha] = \alpha$

Inference:

If $|Z_{cal}| > |Z_{\alpha/2}| (or |Z_\alpha|)$ we reject H_0 otherwise there is no reason to reject H_0.

Problem 4: In a sample of 1000 people in Maharashtra 540 are rice eaters and rest eaters. Can we assume that the rice and wheat equally popular in this state at 1% level of significance.

Solution:

Null hypothesis:

Both rice and wheat are equally popular in the state. $H_0 : p = p_0 = 0.5$

Alternate hypothesis:

Both rice and wheat are not equally popular in the state $H_0 : p_0 \neq 0.5 \, p = 540$

Level of significance:

Test statistic:

$$p = \frac{540}{1000} = 0.540$$

$$n = 1000$$

$$p_0 = 0.5$$

$$Z = \frac{p - p_0}{\sqrt{\frac{p_0 * (1 - p_0)}{n}}}$$

$$= \frac{0.54 - 0.5}{\sqrt{\frac{0.5(0.5)}{1000}}} = \frac{0.04}{0.0158}$$

$$= 2.516$$

$$Z_{tab} = 0.01 \, Z_{\alpha/2} = 2.58$$

$$Z_{cal} = 2.53 < Z_{\alpha/2} = 2.58$$

There is no reason to reject H_0.

Both rice and wheat are equally popular in Maharashtra.

Problem 5: 20 peoples where attacked by a diseases and only 18 survived. Will you reject the hypothesis that the survival rate if attacked by this diseases is 85% in favour of the hypothesis that it is more at 5% level of significance.

Solution:

Null hypothesis:

The survival rate is 85%

Alternative hypothesis:

The survival is more than 85%

Level of significance: $\alpha = 0.05$

Test statistic:

$$p = \frac{18}{20} = 0.9$$
$$n = 20$$
$$p_0 = 0.85$$

$$Z = \frac{p - p_0}{\sqrt{\frac{p_0 * (1 - p_0)}{n}}}$$
$$= \frac{0.9 - 0.85}{\sqrt{\frac{0.85(1 - 0.85)}{20}}} = \frac{0.05}{0.0798}$$
$$= 0.6266$$
$$Z_{tab} = 0.05 \; Z_{\alpha/2} = 1.96$$
$$Z_{cal} = 0.6266 < 1.96 = Z_{\alpha/2}$$

There is no reason to reject H_0

The survival rate is 85%.

TEST FOR DIFFERENCE BETWEEN TWO MEANS:

Null hypothesis:

$H_0 = \mu_1 = \mu_2$ where μ_1 and μ_2 are two population means. In other words,= may be stated as there is no significant difference between two sample means or the two samples have come from the same population.

Alternate hypothesis:

$H_1 : \mu_1 \neq \mu_2$ (Two sided)

$$\left.\begin{array}{l} H_1 : \mu_1 < \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{array}\right\} \text{(One sided)}$$

Level of siginificance:

α is taken to be 0.05 or 0.01 pr it can take a specified lower value.

Test statistic

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - E(\bar{x}_1 - \bar{x}_2)}{S.E\left(\bar{x}_1 - \bar{x}_2\right)}$$

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) \ = \mu - \mu = 0$$

$$S.E(\bar{x}_1 - \bar{x}_2) = \sqrt{v(\bar{x}_1 - \bar{x}_2)}\ 1$$

$$= \sqrt{v(\bar{x}_1) + v(\bar{x}_2)}$$

$$= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

When population variances are unknown then they are replaced by their estimators namely $x_1^2\ s_1^2\ and\ s_2^2$ respectively. Therefore the test statistics becomes

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Here, $\bar{x}_1\ and\ n_1$ refers sample mean and sample size based on the first sample

$\bar{x}_2\ and\ n_2$ refers sample mean and sample size based on the second sample

Critical value:

Depending on alternate hypothesis H_1, the critical values are found using normal table.

α	H_1	Tab.value
0.05	2 sided	1.96
0.01	2 sided	2.58
0.05	1 sided	1.65
0.01	1 sided	2.33

Inference:

If $|Z_{cal}| > Z_{\alpha/2}\ \left(or\ |Z_{cal}| > Z_{\alpha}\right)$

Then the null hypothesis is rejected.

If $|Z_{cal}| < Z_{\alpha/2}$ $\left(or\ |Z_{cal}| < Z_{\alpha}\right)$

Then there is no reason to reject the null hypothesis.

Problem 6: The average hourly wage of a sample of 150 workers in a plant A was 2.56 rupees. 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 the SD of Rs.1.28. Can an applicant safely assume that the hourly wage paid by plant B or higher than those paid by plant A.

Null hypothesis:

The average hourly wage paid by plant A and plant B are same (i.e) $\mu_1 = \mu_2$

Alternate hypothesis:

The average hourly wage paid by plant B is higher than those paid by plant A.

i.e $\mu_1 < \mu_2\,(or)\,\mu_2 > \mu_1$

Test statistic:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$Z = \frac{2.56 - 2.87}{\sqrt{\frac{(1.08)^2}{150} + \frac{(128)^2}{200}}} = \frac{-0.31}{\sqrt{0.0078 + 0.0082}} = \frac{-0.31}{0.1265}$$

Z = -2.4514

$\alpha = 0.05$ $\qquad Z_{tab} = 1.65$

$|Z_{cal}| = 2.4514 > Z_{tab} = 1.65$

The null hypothesis is rejected.

The average hourly wage paid plant B is higher than those paid by A.

TEST FOR SIGNIFICANCE OF DIFFERENCE BETWEEN SAMPLE PROPORTIONS:

Given that two samples if sizes n_1 and n_2 with the proportion p_1 and p_2 respectively. We have to test whether there is any significant difference between p_1 and $p_{2.}$

Null hypothesis:

There is no significant difference between the two sample proportions.

Alternative hypothesis:

There is significant difference between the two sample proportions.

$H_1 : p \neq p_0$

$H_1 : p < p_0$

$H_1 : p > p_0$

Level of significance:

α Is fixed at the level 0.05/0.01

Test statistic and its distribution:

$$Z = \frac{(p_1 - p_2) - E[p_1 - p_2]}{SE(p_1 - p_2)}$$

Under $H_{0,}$ $$Z = \frac{p_1 - p_2}{SE[p_1 - p_2]}$$

$$= \frac{p_1 - p_2}{\sqrt{\frac{p_1 Q_1}{n_1} + \frac{p_2 Q_2}{n_2}}}$$

If population proportion is unknown then it is estimated using

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

That test statistic becomes

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1 - \hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

Critical values:

When $H_1 : p_1 \neq p_2$ from normal table we can find $Z_{\alpha/2}$ using $p[|Z| \geq Z_{\alpha/2}] = \alpha/2$

When $H_1 : p_1 < p_2$ $p_2 (p_1 > p_2)$ from normal table we can find Z_α such that $p[|Z| < Z_\alpha] = \alpha \, (or) \, p[|Z| > Z_\alpha] = \alpha$

Inference:

If H_0 $|Z_{cal}| > Z_{\alpha/2}$ then the null hypothesis is rejected. Otherwise there is no reason to reject H_0.

If $|Z_{cal}| > |Z_{\alpha/2}| (or |Z_\alpha|)$

Problem 7: In a large city A 20% of a random sample of 900 school children have defective eye sight. In other large city B, 15% of a random sample of 1600 children have the same defect. Is this difference between the two proportions significant?

Solution:

Null hypothesis:

There is no significant difference between the two proportions.

Alternative hypothesis:

There is significant difference between the two proportions

Level of significance: $\alpha = 0.05$

$n_1 = 900 \qquad n_2 = 1600$

$p_1 = 0.20 \qquad p_2 = 0.15$

$$\begin{aligned}\hat{p} &= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \\ &= \frac{900(0.2) + 1600(0.15)}{900 + 1600} \\ &= \frac{420}{2500} \\ &= 0.1680\end{aligned}$$

$$\begin{aligned}Z &= \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \\ &= \frac{0.20 - 0.15}{\sqrt{0.1680(0.820)\left[\frac{1}{900} + \frac{1}{1600}\right]}} \\ &= \frac{0.05}{\sqrt{0.0002}}\end{aligned}$$

$Z_{cal} = 3.2200$

$Z_{tab} = 0.05 \qquad Z_{\alpha/2} = 1.965$

$Z_{cal} = 3.22 > 1.965 = Z_{\alpha/2}$

$\therefore$ There is significant difference between the two proportions.

Problem 8: Before an increase in exercise duty on tea 800 persons out of a sample of 1000 persons were found to be tea drinkers. After an excess increase in duty 800 people. Using the above statement check whether there is a significant decrease in the consumption of tea after the increase in excise duty?

Solution:

Null hypothesis:

There is no significant decrease after the increase in excise duty in consumption of tea

Alternative hypothesis:

There is significant decrease in consumption of tea after the increase in excise duty

Level of significance: $\alpha = 0.05$

Test statistic:

$n_1 = 1000 \qquad n_2 = 1200$

$p_1 = 0.8 \qquad p_2 = 0.667$

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$= \frac{100(800) + 1200(800)}{1000 + 1200}$$

$$= \frac{1760000}{2200}$$

$$= 800$$

$$\hat{p} = \frac{1000(0.8) + 1200(0.6667)}{1000 + 1200}$$

$$= 0.7273$$

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

$$= \frac{08 - 0.6667}{\sqrt{0.7273(0.2727)\left[\frac{1}{1000} + \frac{1}{1200}\right]}}$$

$$= \frac{0.1333}{\sqrt{0.0004}} = \frac{0.1333}{0.0191}$$

$Z_{cal} = 6.9905$

$Z_{tab} = 0.05 \qquad Z_{\alpha/2} = 1.65$

$Z_{cal} = 6.9905 > 1.965 = Z_{\alpha}$

Inference: We reject the null hypothesis

There is significant difference in consumption tea after the increase in excise duty.

EXACT TEST/ SMALL SAMPLE TEST

TEST FOR SINGLE MEAN:

Assumption:

The population is normal with mean μ and variance σ^2 .A random sample of size n is drawn from the population.

Population mean $\mu = \mu_0$ is to be tested the other parameter σ^2 may be known or unknown.

Null hypothesis:

There is no significant difference between sample mean and population mean.

Alternative hypothesis:

There is significant difference between the sample mean and population mean

$H_1 : \mu \neq \mu_0$

$H_1 : \mu > \mu_0$

$H_1 : \mu < \mu_0$

Level of significance:

$\alpha = 0.05/0.01$ or any other specified value

Test statistic:

When σ is unknown

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \approx F_{(n-1)}$$

n- sample size

$\overline{X}$ - sample mean

s- unbiased estimation of $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

When σ is known

$$Z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \approx N(0,1)$$

Critical value:

i.) When σ is known

$H_1 : \mu = \mu_0$, using normal tables we can find $Z_{\alpha/2}$ such that $p[|Z| > Z_{\alpha/2}] = \alpha/2$

$H_1 : \mu > \mu_0$ (or) $H_1 : \mu < \mu_0$, using normal tables we can find Z_α such that $p[|Z| > Z_\alpha] = \alpha$ (or) $[p\{|Z| < Z_\alpha\}] = \alpha$

ii.) When is unknown

$H_1 : \mu \neq \mu_0$ using t tables we can find $t_{\alpha/2}$ such that $p\left[|t| > t_{\alpha/2}\right] = \alpha/2$

$H_1 : \mu > \mu_0$ (or) $H_1 : \mu < \mu_0$, using t tables we can find $t_\alpha(n-1)$ such that $p\left[|t| > t_\alpha(n-1)\right] = \alpha$ (or) $p\left[|t| < t_\alpha(n-1)\right] = \alpha$

Inference:

When is σ known, i.) If reject $|Z| > Z_{\alpha/2}$ reject H_0

ii.) If reject $|Z| > Z_{\alpha/2}\,(or\,|Z| < Z_\alpha)$ reject H_0

When is σ unknown i.) For testing $H_1 : \mu \neq \mu_0$ reject if $|t| > t_{\alpha/2}(n-1)$

ii.) For testing $H_1 : \mu < \mu_0\,(or\,\mu > \mu_0)$ reject H_0 if $|t| > t_{\alpha/2}(n-1)$

Problem 9: A random sample of 10 boys have the following IQ values are 70,120,110,101,88,83,95,98,107,100. Do these data support the assumption of the population mean IQ of 100.

Solution:

Null hypothesis:

The population mean IQ is 100

Alternative hypothesis:

The population mean of boys IQ is not 100

$H_1 : \mu \neq 100$

Level of significance: $\alpha = 0.05$

Test statistic:

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \approx t_{(n-1)}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$= \sqrt{\frac{1833.6}{9}}$$

$$= \sqrt{203.7333}$$

$$= 14.2735$$

(n-1)	$t_{\alpha/2}$
8	2.306
9	2.262
10	2.228
11	2.201

$$t = \frac{97.2 - 100}{14.2735 \big/ \sqrt{10}}$$

$$t = 0.6203$$

$$t_{\alpha/2} = 2.262$$

$$|t_{cal}| = 0.6203 < 2.262 = t_{\alpha/2}$$

The population mean of boys IQ is not 100

Problem 10: 10 specimens of copper wire brought from a large lot have the following breaking strength in Kg 578,572,572,568,571,570,570,572,596. Test whether the mean breaking strength of the values may be taken as 578.

Solution:

Null hypothesis:

The mean breaking strength of the value is 578.

Alternative hypothesis:

The mean breaking strength of the value is not 578

Level of significance: $\alpha = 0.05$

Test statistic:

$$t = \frac{\overline{X} - \mu_0}{s / \sqrt{n}} \approx t_{(n-1)}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$= \sqrt{\frac{1212.1}{9}}$$

$$= 11.6051$$

$$t = \frac{571.7 - 578}{11.6051 / \sqrt{10}}$$

$$= -0.1717$$

$$t_{\alpha/2} = 2.262$$

$$|t_{cal}| = 0.1717 < 2.262 = t_{\alpha/2}$$

There is no reason to reject the null hypothesis.

The mean breaking strength of the value is 578

TEST FOR SIGNIFICANCE OF DIFFERENCE BETWEEN TWO MEANS (INDEPENDENT SAMPLES)

Assumptions:

The two population is normal with mean μ and variance σ^2. (i.e) $x_{1,}x_{2,}$.......x_{n1} ~ N(μ_1 , $\sigma_1{}^2$) and y_1,$y_{2,}$.......y_{n2} ~ N(μ_2 , $\sigma_2{}^2$).The population mean μ_1 & μ_2 are unknown. The population variances $\sigma_1{}^2$ & $\sigma_2{}^2$ are equal but unknown (i.e) $\sigma_1{}^2 = \sigma_2{}^2 = \sigma^2$ (unknown). Samples are drawn from the population are independent and random.

Null hypothesis:

There is no significant difference between the two population means (i.e) $H_1 : \mu_1 = \mu_2$

Alternative hypothesis:

There is significant difference between the two population means

$H_1 : \mu_1 \neq \mu_2$

$H_1 : \mu_1 > \mu_2$

$H_1 : \mu_1 < \mu_2$

Level of significance: $\alpha = 0.05 / 0.001$

Test statistic:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - E[\overline{X}_1 - \overline{X}_2]}{SE[\overline{X}_1 - \overline{X}_2]}$$

$$= \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}}$$

Since under $H_0 : \mu_1 = \mu_2$ and σ^2 is unknown, σ^2 is replaced by its estimation

n_1 - first sample size

n_2 - second sample size

$\overline{X}_1$ - first sample mean

$\overline{X}_2$ - second sample mean

$$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

s_1^2 - first sample variance

s_2^2 - second sample variance

$$\therefore t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{n_1 + n_2}{n_1 n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

Critical value:

For two sided test, we can find $t_{\alpha/2}(n_1 + n_2 - 2)$ from t tables for $(n_1 + n_2 - 2)$ degree of freedom. For one sided test, we can find $t_{\alpha}(n_1 + n_2 - 2)$ from t tables for $(n_1 + n_2 - 2)$ degree of freedom

Inference:

For two sided test, reject H_0 if $|t_{cal}| > t_{\alpha/2}(n_1 + n_2 - 2)$ otherwise there is no reason to reject H_0

For one sided test, reject H_0 if $|t_{cal}| > t_{\alpha}(n_1 + n_2 - 2)$ otherwise there is no reason to reject H_0

Problem 11: The heights of six randomly chosen sailors are (in inches) 63,65,68,69,71,72,73. Discuss the light that these data throw on the suggestion that the sailors are on the average taller than soldiers.

Solution:

Null hypothesis:

There is no significant difference between the height of the sailors and soldiers $H_0 : \mu_1 = \mu_2$

Alternative hypothesis:

There is significant difference between the height of the sailors and soldiers $H_1 : \mu_1 > \mu_2$

Level of significance: $\alpha = 0.05$

Test statistics:

$n_1 = 6$ $\qquad n_2 = 10$

$\overline{X}_1 = 68$ $\qquad \overline{X}_2 = 67.8$

$$s_1^{\ 2} = \frac{1}{6}(60) = 10$$

$$s_2^{\ 2} = \frac{1}{10}(153.6) = 15.36$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{n_1 + n_2}{n_1 n_2}\right)}}$$

$$= \frac{68 - 67.8}{\sqrt{\frac{69(10) + 10(15.36)}{6 + 10 - 2}\left(\frac{6 + 10}{60}\right)}}$$

$$= 0.0992$$

$d.f - n_1 + n_2 - 2 = 14$

$t_{14(\alpha)} = 0.0992 < 1.761 = t_\alpha (14)$

$\therefore$ There is no reason to reject H_0

$\therefore$ The sailors are on the average taller the soldiers

TEST FOR DIFFERENCE BETWEEN MEANS PAIRED VALUES (DEPENDENT SAMPLE TEST/ PAIRED T-TEST)

Assumption:

The population is normal with mean μ and variance σ^2

A random sample of observations $(x_1, y_1), (x_2, y_2)\ldots..(x_n, y_n)$ is drawn from the population

μ_d and $\sigma_d{}^2$ are unknown

$\mu_d = \mu_0$ is to be tested

Null hypothesis:

There is no significant difference between the means

Alternative hypothesis:

There is significant difference between the means

$H_1 : \mu_d \neq \mu_0$

$H_1 : \mu_d > \mu_0$

$H_1 : \mu_d < \mu_0$

Level of significance: $\alpha = 0.05 / 0.01 / 0.001$

Test statistic:

$$t = \frac{\bar{d} - E(\bar{d})}{SE(\bar{d})} \quad where\ \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

$$under\ H_0, E(\bar{d}) = \mu_0$$

$$SE(\bar{d}) = \frac{\hat{\sigma}_d}{\sqrt{n-1}}$$

$$\hat{\sigma}_d = \sqrt{\frac{1}{n}\sum d_i^2 - \bar{d}^2}$$

$$t = \frac{\bar{d}}{\hat{\sigma}_d / \sqrt{n-1}} \sim t_{(n-1)}$$

Critical value:

For 2 sided test, we find $t_{\alpha/2} = (n-1)$from t-table for n-1 degree of freedom

For 1 sided test, we find $t_{\alpha} = (n-1)$from t-table for n-1 degree of freedom

Inference:

If $|t_{cal}| > t_{\alpha/2} = (n-1)$ then reject $H_0$for 2 sided test.

Reject H_0 if $|t_{cal}| > t_{\alpha} = (n-1)$ for 1sided test

Problem 12: Eleven school boys were given a test in statistics, they were given one month tuition and then second test was conducted. The marks obtained by them in the first and second tests are given below. Do the marks give the evidences that the students are benefitted by extra coaching.

Marks in 1st test	Marks in 2nd test	Marks in 1st test	Marks in 2nd test
23	24	17	20
20	19	23	23
19	22	16	20
21	18	19	18
28	20		
20	22		
18	20		

Solution:

Null hypothesis:

The students are not benefited by the extra coaching.

Alternative hypothesis:

The students are benefited by the extra coaching.

Level of significance: $\alpha = 0.05$

Test statistic:

$$t = \frac{\bar{d}}{\hat{\sigma}_d / \sqrt{n-1}} \sim t_{(n-1)}$$

x_i	y_i	$d_i = x_i - y_i$	d_i^2
23	24	-1	1
20	19	1	1
19	22	-3	9
21	18	3	9
28	20	8	64
20	22	-2	4
18	20	-2	4
17	20	-3	9
23	23	0	0
16	20	-4	16
19	18	1	1

$$\begin{aligned}\hat{\sigma}_d &= \sqrt{\frac{1}{n}\sum d_i^2 - \bar{d}^2} \\ &= \sqrt{\frac{1}{11}(118) - 0.0331} \\ &= 3.2702\end{aligned}$$

$$\begin{aligned}t &= \frac{-0.1818}{3.2708/\sqrt{10}} \\ &= \frac{-0.1818}{1.0341} = -0.1758\end{aligned}$$

Degree of freedom, n-1=11-1=10

$$t_\alpha^{10} = 1.81$$

$$\therefore \left|t_{cal}\right| = 0.1758 < 1.81 = t_\alpha^{10}$$

There is no reason to reject H_0 .

The students are not benefitted by the extra coaching.

Problem 13: The scenes of 10 candidates prior and after training are given below.

Prior training	84	48	36	37	54	69	83	95	90	1
After training	90	58	56	49	62	81	84	86	84	7

Is the training is effective?

Solution:

Null hypothesis:

The training is not effective

Alternative hypothesis:

The training is effective

Level of significance: $\alpha = 0.05$

Test statistic:

x_i	y_i	$d_i = x_i - y_i$	d_i^2
84	90	-6	36
48	58	-10	100
36	56	-20	400
37	49	-12	144
54	62	-8	64
69	81	-12	144
8	84	-1	1
96	86	10	100
90	84	6	36
65	75	-10	100

$$\hat{\sigma}_d = \sqrt{\frac{1}{n}\sum d_i^2 - \bar{d}^2}$$

$$= \sqrt{\frac{1}{10}(1125) - 39.69}$$

$$= 8.5329$$

$$t = \frac{-6.3}{8.5329/\sqrt{9}}$$

$$t_{cal} = 2.2150$$

Degrees of freedom = n-1= 9

$$t_\alpha = 1.83$$

$$t_{cal} = 2.2150 > t_{tab} = 1.83$$

We reject the null hypothesis and we conclude that the training is effective.

One-Way ANOVA

A One-Way Analysis of Variance is a way to test the equality of three or more means at one time by using variances.

Assumptions

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent.
- The variances of the populations must be equal.

Hypotheses

- The null hypothesis will be that all population means are equal, the alternative hypothesis is that at least one mean is different.
- In the following, lower case letters apply to the individual samples and capital letters apply to the entire set collectively. That is, n is one of many sample sizes, but N is the total sample size.

Grand Mean

The grand mean of a set of samples is the total of all the data values divided by the total sample size. This requires that you have all of the sample data available to you, which is usually the case, but not always. It turns out that all that is necessary to find perform a one-way analysis of variance are the number of samples, the sample means, the sample variances, and the sample sizes.

$$\overline{X}_{GM} = \frac{\sum x}{N}$$

Another way to find the grand mean is to find the weighted average of the sample means. The weight applied is the sample size.

$$\overline{X}_{GM} = \frac{\sum n\,\overline{x}}{\sum n}$$

Total Variation

The total variation (not variance) is comprised the sum of the squares of the differences of each mean with the grand mean.

$$SS(T) = \sum\left(x - \overline{X}_{GM}\right)^2$$

There is the between group variation and the within group variation. The whole idea behind the analysis of variance is to compare the ratio of between group variance to within group variance. If the variance caused by the interaction between the samples is much larger when compared to the variance that appears within each group, then it is because the means aren't the same.

Between Group Variation

The variation due to the interaction between the samples is denoted SS(B) for Sum of Squares Between groups. If the sample means are close to each other (and therefore the Grand Mean) this will be small.

$$SS(B) = \sum n\left(\bar{x} - \bar{X}_{GM}\right)^2$$

There are k samples involved with one data value for each sample (the sample mean), so there are k-1 degrees of freedom.

The variance due to the interaction between the samples is denoted MS(B) for Mean Square Between groups. This is the between group variation divided by its degrees of freedom. It is also denoted by s_b^2.

Within Group Variation

The variation due to differences within individual samples, denoted SS(W) for Sum of Squares Within groups. Each sample is considered independently, no interaction between samples is involved. The degrees of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: df = N - k.

$$SS(W) = \sum df \ .s^2$$

The variance due to the differences within individual samples is denoted MS(W) for Mean Square Within groups. This is the within group variation divided by its degrees of freedom. It is also denoted by s_w^2. It is the weighted average of the variances (weighted with the degrees of freedom).

F test statistic

ANOVA Test Statistic Recall that a F variable is the ratio of two independent chi-square variables divided by their respective degrees of freedom. Also recall that the F test statistic is the ratio of two sample variances, well, it turns out that's exactly what we have here. The F test statistic is found by dividing the between group variance by the within group variance. The degrees of freedom for the numerator are the degrees of freedom for the between group (k-1) and the degrees of freedom for the denominator are the degrees of freedom for the within group (N-k).

$$F = \frac{s_b^2}{s_w^2}$$

Summary Table

	SS	df	MS	F
Between	SS(B)	k-1	SS(B)/k-1	MS(B)/MS(W)
Within	SS(W)	N-k	SS(W)/N-k	
Total	SS(T)=SS(B)+SS(W)	N-1		

Notice that each Mean Square is just the Sum of Squares divided by its degrees of freedom, and the F value is the ratio of the mean squares. Do not put the largest variance in the numerator, always divide the between variance by the within variance. If the between variance is smaller than the within variance, then the means are really close to each other and you will fail to reject the claim that they are all equal. The degrees of freedom of the F-test are in the same order they appear in the table

Decision Rule

The decision will be to reject the null hypothesis if the test statistic from the table is greater than the F critical value with k-1 numerator and N-k denominator degrees of freedom.

If the decision is to reject the null, then at least one of the means is different. However, the ANOVA does not tell you where the difference lies. For this, you need another test, either the Scheffe' or Tukey test.

Problem 14: Four models of lacrosse helmets were compared. Measurements of Gadd severity index were made on each of 10 hits per helmet. Test whether helmet means are significantly different at α =0.05 significance level.

Brand	**Mean**	**SD**	**Sample Size**
Sports Helmets Cascade	1166.1	152.40	10
Sports Helmets Cascade Air Fit	1117.6	216.23	10
Sports Helmets Ultralite	857.0	151.54	10
Bacharach Ultralite	1222.8	123.08	10

Solution:

Hypothesis: There is no significance difference between lacrosse helmets

Alternative Hypothesis: There is significance difference between lacrosse helmets

Level of Significance: α =0.05

Test Statistic:

Source	SS	df	MS	F
Between	784747.5	3	261582.5	9.68
Within	972848.1	36	27023.56	
Total	1757595.6	39		

Rejection Region: Fobs ≥ $F_{.05,3,37}$ = 2.88 (appox)

We reject the null hypothesis. Therefore, There is significance difference between lacrosse helmets.

Problem 15: The fog index measures the reading difficulty based on the average number of words pe sentence and percent of words with 3 or more syllables. High values of the fog index are associated with difficult reading levels. Independent random samples of six ads were taken from 3 magazines. Test for “magazine effects” based on the F-test for 5% level of significance.

Scientific American: 11.16, 9.23, 15.75, 8.20, 9.92, 11.55

Fortune : 12.63, 9.42, 9.87, 11.46, 10.77, 9.93

New Yorker: 8.15, 6.37, 8.28, 6.37, 5.66, 9.27

Solution:

Hypothesis: There is no significance difference between 3 magazines

Alternative Hypothesis: There is significance difference between 3 magazines

Level of Significance: α =0.05

Test Statistic:

Source	SS	df	MS	F
Between	48.53	2	24.27	6.97
Within	52.21	15	3.48	
Total	100.74	17		

Rejection Region: $F_{.05,2,15}$=3.68<F_{obs}=6.97

We reject the null hypothesis. Therefore, there is significance difference between 3 magazines.

Bartlett’s Test for Homogeneity of Variance:

Bartlett's test (Snedecor and Cochran, 1983) is used to test if k samples have equal variances. Equal variances across samples is called homogeneity of variances. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Bartlett test can be used to verify that assumption.

Bartlett's test is sensitive to departures from normality. That is, if your samples come from non-normal distributions, then Bartlett's test may simply be testing for non-normality. The Levene test is an alternative to the Bartlett test that is less sensitive to departures from normality.

Procedure:

Null Hypothesis: There is no difference between k sample variances

ie., $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$

Alternative Hypothesis: There is difference between any two sample variances

ie., $H_1 : \sigma_i^2 \neq \sigma_j^2$

Level of significance: $\alpha = 0.05/0.01/0.001$

Test statistic

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(N_i - 1)\ln s_i^2}{1 + \left(1/(3(k-1))\right)\left(\left(\sum_{i=1}^{k} 1/(N_i - 1)\right) - 1/(N-k)\right)}$$

In the above, s_i^2 is the variance of the ith group, N is the total sample size, N_i is the sample size of the i th group, k is the number of groups, and s_p^2 is the pooled variance. The pooled variance is a weighted average of the group variances and is defined as:

$$s_p^2 = \sum(N_i - 1)s_i^2 / (N - k)$$

Critical Region: The variances are judged to be unequal if, $T > \chi^2_{1-\alpha,k-1}$ where $\chi^2_{1-\alpha,k-1}$ is the critical value of the chi-square distribution with k - 1 degrees of freedom and a significance level of α.

TEST FOR SIGNIFICANCE OF CORRELATION COEFFICIENT:

Assumption:

- the population is bivariate normal population.
- The population correlation coefficient is assumed to be zero.
- A random sample of size n drawn from the population and the sample correlation coefficient is taken as r.

Null hypothesis: there is no significant difference in the correlation coefficient i.e correlation coefficient in the population is assumed to be zero. ie., $H_0 : \rho = 0$

Alternative hypothesis: $H_1 : \rho \neq 0, H_1 : \rho > 0, H_1 : \rho < 0$

Level of significance: $\alpha = 0.05 \; or \; 0.01$

Test statistic:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{(n-2)}$$

Critical value: from t table, we can find $t_{\frac{\alpha}{2},(n-2)}$ for the given α and n-2 degrees of freedom.

Inference: for the two sided test reject $H_0 \; if \; t_{cal} > t_{\frac{\alpha}{2},(n-2)}$ otherwise there is no reason to reject H_0 .

Problem 15: A random sample of 27 pairs observation from a bivariate normal population give a correlation coefficient of 0.42 can you conclude that the variables in the population are uncorrelated.

Solution:

Null hypothesis: the variables in the population are uncorrelated $H_0 : \rho = 0$

Alternative hypothesis: the variables in the population are correlated $H_1 : \rho \neq 0$

Level of significance: $\alpha = 0.05$

Test statistic:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{(n-2)}$$

$$t = \frac{0.36}{\sqrt{\frac{1-0.36^2}{17-2}}}$$

$$t = \frac{0.36}{0.2409}$$

$$t = 1.4944$$

Degrees of freedom = 17-2= 15

$$t_\alpha(15) = 1.75$$

$$t_{cal} = 1.4944 < t_{tab} = 1.75$$

We reject the null hypothesis and we conclude that the variables are uncorrelated.

TEST FOR SIGNIFICANCE OF REGRESSION COEFFICIENT:

Assumption:

1. The population is bivariate normal with regression coefficient of Y on X is β
2. β is unknown.
3. A random sample of size n is drawn from bivariate normal population and its regression coefficient of Y on X is b.

Null hypothesis: there is no significant difference between sample regression coefficient and population regression coefficient.

Alternative hypothesis: there is significant difference between sample regression coefficient and population regression coefficient. $H_1 : \beta \neq \beta_0, H_1 : \beta > \beta_0, H_1 : \beta < \beta_0$

Level of significance: $\alpha = 0.05$ Test statistics:

$$t = \frac{b - E(b)}{S.E(b)}$$

$$t = \frac{b - \beta_0}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}} \sim t_{(n-2)}$$

Where x,y are the sample observation n- sample size

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

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^{\ 2} - (\bar{x})^2}$$

$$\hat{y} = \hat{a} + \hat{b}x_i$$
$$a = \bar{y} - b\bar{x}$$

Critical value: from t table, we can find $t_{\frac{\alpha}{2},(n-2)}$ from t table for n-2 degrees of freedom.

For one sided test, we can find $t_{\alpha,(n-2)}$ from t table for n-2 degrees of freedom.

Inference: for the two sided test reject $H_0\ if\ t_{cal} > t_{\frac{\alpha}{2},(n-2)}$ otherwise there is no reason to reject H_0 .

For the one sided test reject $H_0\ if\ t_{cal} > t_{\alpha,(n-2)}$ otherwise there is no reason to reject H_0 .

Problem 16: Test the significance of regression coefficient by X if the following are the values of sample drawn from bivariate normal population.

X	1	2	3	4	5	6
Y	10	12	14	16	14	15

Solution:

Null hypothesis: the regression equation is linear. $H_0 : \beta = 0$

Alternative hypothesis: the regression equation is linear. $H_1 : \beta \neq 0$

Level of significance: $\alpha = 0.05$

Test statistic: assume $\beta = 0$

x_i	y_i	$x_i y_i$	$x_i - \bar{x}$	$x_i^{\ 2}$	$\hat{y}_i$	$(y_i - \bar{y}_i)^2$	$(x_i - \bar{x})^2$
1	10	10	-2.5	1	11.143	1.3064	6.25
2	12	24	-1.5	4	121.0858	0.0074	2.25
3	14	42	-0.5	9	13.0286	0.9436	0.25
4	16	64	0.5	16	13.9714	4.1152	0.25
5	14	70	1.5	25	14.9142	0.8358	2.25
6	15	90	2.5	36	15.857	0.7344	6.25

$\bar{x} = 3.5$

$\bar{y} = 13.5$

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^{\ 2} - (\bar{x})^2}$$

$$b = \frac{50 - 47.25}{15.1667 - 12.25}$$

$b = 0.9428$

$a = \bar{y} - b\bar{x}$

$a = 13.5 - (0.9428)(3.5)$

$a = 10.2002$

$$t = \frac{b - \beta_0}{\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}}$$

$$t = \frac{0.9428 - 0}{\sqrt{\dfrac{7.9428}{4(17.5)}}}$$

$t_{cal} = 2.7985$

$t_{\frac{\alpha}{2}}(4) = 2.776$

$t_{cal} = 2.7985 > t_{\frac{\alpha}{2}}(4) = 2.776$

We reject null hypothesis. Therefore, the regression equation is linear.

TEST FOR SIGNIFICANCE OF PARTIAL CORRELATION COEFFICIENT:

Assumption:

1. The population is multivariate normal with partial correlation coefficient ρ of order k.
2. A random sample is drawn from a population with the sample partial coefficient coefficient r of order k.

Null hypothesis: the population partial coefficient ρ of order k is not significant $H_0 : \rho = 0$

Alternative hypothesis: $H_1 : \rho \neq 0, H_1 : \rho > 0, H_1 : \rho < 0$

Level of significance: $\alpha = 0.05\, or\, 0.01$

Test statistic:

$$t = \frac{r - E(r)}{SE(r)}$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-k-2}}} \sim t_{(n-2)}$$

Where, n- sample size, k-order of partial correlation coefficient, r-partial correlation coefficient of the samples.

Critical value: from t table, we can find $t_{\frac{\alpha}{2},(n-k-2)}$ from t table for n-k-2 degrees of freedom.

For one sided test, we can find $t_{\alpha,(n-k-2)}$ from t table for n-k-2 degrees of freedom.

Inference: for the two sided test reject $H_0\, if\, t_{cal} > t_{\frac{\alpha}{2},(n-k-2)}$ otherwise there is no reason to reject H_0 .

For the one sided test reject $H_0\, if\, t_{cal} > t_{\alpha,(n-k-2)}$ otherwise there is no reason to reject H_0 .

Problem 17: A sample of size 10 observation from trivariate normal population gave the partial correlation coefficient between first and second variable as 0.3247 is this significant at 5% level.

Solution:

Null hypothesis: the partial correlation coefficient of order 1 is not significant $H_0 : \rho_{12.3} = 0$

Alternative hypothesis: the partial correlation coefficient of order 1 is significant $H_1 : \rho_{12.3} \neq 0$

Level of significance: $\alpha = 0.05$

Test statistic:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-k-2}}} \sim t_{(n-2)}$$

$$t = \frac{0.3247}{\sqrt{\frac{1-(0.3247)^2}{10-1-2}}} = \frac{0.3247}{\sqrt{0.1278}} = 0.9083$$

$$t_{\frac{\alpha}{2},(n-k-2)} = 2.365$$

$$t_{cal} = 0.9083 < t_{\frac{\alpha}{2},(7)} = 2.365$$

We accept null hypothesis and we conclude that the partial correlation coefficient of order 1is not significant.

Show that the test statistic for testing mean of a normal population with unknown $\sigma \sim t_{(n-1)}$

Test statistic:

$$t = \frac{r - E(r)}{SE(r)}$$

$$t = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

$$under\, H_0 : t = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$$

$$where \;\; \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\chi^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$t = \frac{z}{\sqrt{\frac{\chi^2}{n-1}}}$$

$$t = \frac{\overline{X} - \mu_0 / \sigma / \sqrt{n}}{\sqrt{\frac{ns^2}{n-1}}} = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{\sigma} .\sqrt{\sigma^2 (n-1)/n.s^2} = \frac{\overline{X} - \mu_0}{s / \sqrt{n-1}} \sim t(n-1)$$

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

EXACT TEST BASED ON F- DISTRIBUTION:

Test for ratio of two variance or test for equality of two variances:

Assumption:

1. Two independent normal population $N\left(\mu_1, \sigma_1^2\right)$ and $N\left(\mu_2, \sigma_2^2\right)$ are considered.
2. Two samples are drawn from the given population let n_1 be the size of the first sample and n_2 be the size of the second sample with the sampling variance s_1^2 and s_2^2.
3. Variances of the population are unknown and assumed to be equal.

Null hypothesis: the population variances are equal $H_0 : \sigma_1^2 = \sigma_2^2$

Alternative hypothesis: the population variance not equal. $H_1 : \sigma_1^2 \neq \sigma_2^2, H_1 : \sigma_1^2 > .\sigma_2^2, H_1 : \sigma_1^2 < \sigma_2^2$

Level of significance: $\alpha = 0.05 \, or \, 0.01$

Test statistic:

$$F_{cal} = \frac{n_1 s_1^2 / \sigma_1^2 (n_1 - 1)}{n_2 s_2^2 / \sigma_2^2 (n_2 - 1)} \sim F(n_1 - 1, n_2 - 1)$$

$$F_{cal} = \frac{n_1 s_1^2 / (n_1 - 1)}{n_2 s_2^2 / (n_2 - 1)}$$

$$F_{cal} = \frac{n_1 s_1^2 / (n_2 - 1)}{n_2 s_2^2 / (n_1 - 1)}$$

Let $S_1^2 = \frac{n_1}{n_1 - 1} s_1^2 \qquad S_2^2 = \frac{n_1}{n_1 - 1} s_2^2$

$F_{cal} = \frac{S_1^2}{S_2^2}$ where S_1^2 and S_2^2 are unbiased estimators of σ_1^2 and σ_2^2 respectively.

Critical region:

1. For two sided test, we find F_1 and F_2 from F table for $F\ ((n_1 - 1),(n_2 - 1))$ degrees of freedom.
2. When alternative is $H_1 : \sigma_1^2 > .\sigma_2^2$ we can find from F table $F_\alpha((n_1 - 1),(n_2 - 1))$ for $(n_1 - 1),(n_2 - 1)$ degrees of freedom.
3. When alternative is $H_1 : \sigma_1^2 < .\sigma_2^2$ we can find from F table $F_\alpha((n_1 - 1),(n_2 - 1))$ for $(n_1 - 1),(n_2 - 1)$ degrees of freedom.

Inference:

1. For two sided test, if $F_{cal} > F_{\frac{\alpha}{2}}(n_1 - 1),(n_2 - 1)$ then we reject H_0 otherwise there is no reason to reject H_0
2. For one sided test, $H_1 : \sigma_1^2 > .\sigma_2^2 \quad F_{cal} > F_\alpha(n_1 - 1),(n_2 - 1)$ then we reject H_0 otherwise there is no reason to reject H_0
3. For one sided test, $H_1 : \sigma_1^2 < .\sigma_2^2 \quad if\ F_{cal} > F_{1-\alpha}(n_1 - 1),(n_2 - 1)$ then we reject H_0 otherwise there is no reason to reject H_0

Problem 18: the following are the values (in 1000's) of an inch obtained by 2 engineers with 10 successive measurement in the same micrometer. Is one engineer significantly more consistent than the other?

Engineer A	503	505	497	505	495	502	499	493	510	501
Engineer B	502	497	492	498	499	495	497	496	498	

Solution:

Null hypothesis: there is no significant difference between consistency of engineers $H_0 : \sigma_1^2 = \sigma_2^2$

Alternative hypothesis: one engineer is more consistent than the other $H_0 : \sigma_1^2 > \sigma_2^2$.

Level of significance: $\alpha = 0.05$

Test statistic:

x_1	x_2	$(x_{1i} - \bar{x}_1)^2$	$(x_{2i} - \bar{x}_2)^2$
503	502	4	23.9013
505	497	16	0.0123
497	492	16	26.1233

505	498	16	0.7901
495	499	36	3.5679
502	495	1	4.4567
499	497	4	0.0123
493	496	64	1.2345
510	498	81	0.7901
501		0	

$$\bar{x}_1 = \frac{5010}{10} = 501$$

$$\bar{x}_2 = \frac{4474}{9} = 479.1$$

$$s_1^2 = \frac{1}{9}(238) = 26.4444$$

$$s_2^2 = \frac{1}{8}(60.8885) = 7.6111$$

$$s_1^2 > s_2^2$$

$$F_{cal} = \frac{26.4444}{7.6111} = 3.4745$$

$$F(9,8) = 3.4$$

$$F_{cal} = 3.4575 > F(9,8) = 3.4$$

We reject the null hypothesis and therefore one engineer is more consistent than the other.

TEST FOR SIGNIFICANCE OF MULTIPLE CORRELATION COEFFICIENT:

Assumption:

1. The population is multivariate normal
2. A random sample of size n is drawn from the population and desired multiple correlation coefficient is obtained $r_{1.234\cdots k+1}$
3. We want to test whether there exist multiple correlation in the population $R_{1.234\cdots k+1}$

Null hypothesis: the multiple correlation in the population is zero $R_{1.234\cdots k+1} = 0$

Alternative hypothesis: $R_{1.234\cdots k+1} > 0$

Level of significance: $\alpha = 0.05 \; or \; 0.01$

Test statistic:

$$F = \frac{r^2}{1-r^2}.\frac{n-k-1}{k} \sim F_{(k,n-k-1)}$$

Where, n- sample size, k-order of multiple correlation coefficient

Critical value: we find $F_\alpha(k,n-k-1)$ from F take for $(k,n-k-1)$ degrees of freedom.

Inference: if $F_{cal} > F_\alpha(k,n-k-1)$then we reject null hypothesis, otherwise there is no reason to reject null hypothesis.

Problem 19: from a 5 variate normal population a random sample of size 20 is taken and multiple correlation coefficient $r_{1.2345}$ is found to be 0.27 test at 5% level the existence of multiple correlation coefficient in the population.

Solution:

Null hypothesis: there does not exist population multiple correlation coefficient $H_0 : r_{1.234\cdots k+1} = 0$

Alternative hypothesis: there exist population multiple correlation coefficient $H_1 : r_{1.234\cdots k+1} = 0$

Level of significance: $\alpha = 0.05$

Test statistic:

$$F = \frac{r^2}{1-r^2}.\frac{n-k-1}{k}$$

$$F = \frac{(0.27)^2}{1-(0.27)^2}.\frac{20-4-1}{4}$$

$$F = \frac{1.0935}{3.7084}$$

$$F = 0.2949$$

n-k-1=15 $t_\alpha(4,15) = 3.06$

$$t_{cal} = 0.2949 < t_\alpha(4,15) = 3.06$$

We accept null hypothesis. There does not exist population multiple correlation coefficient.

EXACT TEST BASED ON CHI-SQUARE:

TEST FOR SIGNIFICANCE OF VARIANCE:

Assumption:

1. The population is normal population mean μ and variance σ^2
2. μ and σ^2 are unknown
3. A random sample of size n is drawn from normal population with mean μ and variance σ^2

Null hypothesis: there is no significant difference between sample variance and population variance. $H_0 : \sigma_1 = \sigma_2$

Alternative hypothesis: there is significant difference between sample variance and population variance $H_1 : \sigma_1 \neq \sigma_2 , H_1 : \sigma_1 > .\sigma_2 , H_1 : \sigma_1 < \sigma_2$

Level of significance: $\alpha = 0.05 \, or \, 0.01$

Test statistic:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_0} \right)^2$$

$$\chi^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sigma_0{}^2}$$

$$= \frac{ns^2}{\sigma_0^2} \sim \chi^2_{(n-1)}..$$

$n - sample\,size$

$s^2 - sample\ \mathrm{var}\,iance$

$\sigma_0^2 - poplation\,\mathrm{var}\,iance\,under\;H_0$

Critical value:

1. For two sided test, we find χ^2 values such that $\chi^2 = \chi^2_{\frac{\alpha}{2}}$ for (n-1) degrees of freedom and $\chi^2 = \chi^2_{1-\frac{\alpha}{2}}$ for (n-1) degrees of freedom.
2. For one sided test, we find χ^2 values such that χ^2 for (n-1) degrees of freedom and α for level of significance.
3. For one sided test, we find $H_0 : \sigma^2 < \sigma_0^2$ values such that $\chi^2_{(1-\alpha)}$ for (n-1) degrees of freedom and χ^2 for (n-1) degrees of freedom.

4. Inference: if $\chi^2_{cal} > \chi^2_{\alpha/2}(n-1)\, or\, \chi^2_{cal} > \chi^2_{\alpha}(n-1)$ we reject null hypothesis there is no reason to reject null hypothesis.

Problem 20: A random sample size of size 10 is taken from normal population and the observation are 2.3 2.4 2.5, 2.7, 2.5, 2.6, 2.6, 2.7, 2.5, 2.4. test the hypothesis that the population variance is 0.16 against the alternative the population variance is greater than 0.16.
Solution:

Null hypothesis: the population variance is 0.16.
Alternative hypothesis: the population variance is greater than 0.16
Level of significance: $\alpha = 0.05$
Test statistic:

x	$(x_i - \bar{x}_1)^2$
2.3	0.0484
2.4	0.0144
2.5	0.0004
2.7	0.0324
2.5	0.0004
2.6	0.0064
2.6	0.0064
2.7	0.0324
2.5	0.0004
2.4	0.0144

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x}_1)^2$$
$$= \frac{0.156}{10}$$
$$= 0.0156$$

$$\chi^2 = \frac{ns^2}{\sigma_0^2}$$
$$= \frac{10(0.0156)}{0.16}$$
$$= 0.975$$

n-1=10-1=9 $\chi_{\alpha}^{2}(9)=16.92$ $\chi_{cal}^{2}=0.975<\chi_{\alpha}^{2}(9)=16.92$

we accept the null hypothesis and therefore the population variance is 0.16.

TEST FOR SIGNIFICANCE INDEPENDENCE OF ATTRIBUTES ASSOCIATION OF ATTRIBUTES:

Contingency table:

It is a two way table for attributes different levels of two attributes are considered and the table gives frequencies corresponding to ith level of one attributes and jth level of an other attribute i=1,2,..,m and j=1,2,..n. this type of table is called contingency table of m*n. for example wwe have the following 2X5 contingency table.

Sex	Illiterate	School education	College edu. Non professional	Professional education	Others
Male	20	15	25	10	5
Female	15	25	20	8	7

Null hypothesis: there is no association between the two attributes (or) the two attributes A and B arer as independent.

Alternative hypothesis: there is association between the two attributes A and B

Level of significance: $\alpha = 0.05 \, or \, 0.01 \, or \, any \, other \, specified \, value$

Test statistic:

$$\chi_{cal}^{2}=\sum_{i=1}^{m}\sum_{j=i}^{n}\left(\frac{\left(O_{ij}-E_{ij}\right)^{2}}{E_{ij}}\right)\sim\chi_{(m-1)(n-1)}^{2}$$

Where O_{ij} is the observed frequency for (i,j)th cell in the contingency table.

Where E_{ij} is the expected frequency for (i,j)th cell in the contingency table and is given by,

$$C_{ij}=\frac{R_iC_j}{N}=E_{ij}$$

Where R_i is the total of ith row

C_j is the total of jth column

N is the grand total.

Critical value: we can find $\chi_{(m-1)(n-1)}^{2}$ from χ^{2} for (m-1)(n-1) degrees of the freedom at level of significance.

Inference : $\chi^2_{cal} > \chi^2_{\alpha,(m-1)(n-1)}$ we reject null hypothesis otherwise there is no reason to reject null hypothesis.

Problem 21: The following table is collected on two characters.

	Cine goers	Non-cine goers
Illiterate	45	68
Literate	83	57

Based on this can you conclude that there is no relation between the habit of cinema going and literacy.

Solution:

Null hypothesis: there is no relation between the habit of cinema going and literacy.

Alternative hypothesis: there is relation between the habit of cinema going and literacy.

Level of significance: $\alpha = 0.05$

Test statistic:

$E_{11} = \frac{(128)113}{253} = 57.17$ $E_{12} = \frac{(125)113}{253} = 55.83$ $E_{21} = \frac{(128)140}{253} = 70.83$

$E_{22} = \frac{(125)140}{253} = 69.17$

O_{ij}	E_{ij}	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
45	57.17	148.1089	2.5907
83	70.83	148.1089	2.0910
68	55..83	148.1089	2.653
57	69.17	148.1089	2.1412

$\chi^2_{cal} = 9.4759$

$\chi^2_{tab} = 3.841$

Inference: $\chi^2_{cal} = 9.4759 > \chi^2_{tab} = 3.841$ we reject the null hypothesis. There is no relation between the habit of cinema going and literacy.

Theorem 1: Show that 2X2 contingency table with frequencies a,b,c,and d χ^2 statistic is $\frac{N(ab-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ where N=a+b+c+d.

Proof:

The observed frequencies are $O_{11} = a \quad O_{12} = b \quad O_{21} = c \quad O_{22} = d$ the expected frequencies are.

$$\frac{R_i C_j}{N} E_{ij}$$

Where R_i is the total of ith row

C_j is the total of jth column

N is the grand total.

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=i}^{n}\left(\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right)$$

$$E_{11} = \frac{(a+b)(a+c)}{N} \qquad E_{12} = \frac{(a+b)(b+d)}{N}$$

$$(O_{11}-E_{11})^2 = \left[a - \frac{(a+b)(a+c)}{N}\right]^2$$

$$(O_{11}-E_{11})^2 = \left[\frac{a(a+b+c+d)-(a+b)(a+c)}{N}\right]^2$$

$$\frac{(O_{11}-E_{11})^2}{E_{11}} = \left[\frac{(ad-bc)}{N}\right]^2$$

$$\frac{(O_{11}-E_{11})^2}{E_{11}} = \frac{(ad-bc)^2.N}{N^2(a+b)(a+c)}$$

$$\frac{(O_{11}-E_{11})^2}{E_{11}} = \frac{(ad-bc)^2}{N\ (a+b)(a+c)}$$

Similarly,

$$\frac{(O_{12}-E_{12})^2}{E_{12}} = \frac{(ad-bc)^2}{N\ (a+b)(b+d)}$$

$$\frac{(O_{21}-E_{21})^2}{E_{21}} = \frac{(ad-bc)^2}{N\ (a+c)(c+d)}$$

$$\frac{(O_{22}-E_{22})^2}{E_{22}} = \frac{(ad-bc)^2}{N\ (b+d)(c+d)}$$

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=i}^{n}\left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right)$$

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

$$\chi^2 = \frac{(ad-bc)^2}{N}\left[\frac{(b+d)(c+d)+(a+c)(c+d)+(a+b)(b+d)+(a+b)(a+c)}{(a+b)(a+c)(b+d)(c+d)}\right]$$

$$\chi^2 = \frac{(ad-bc)^2}{N}\left[\frac{(a+b+c+d)(c+d+a+b)}{(a+b)(a+c)(b+d)(c+d)}\right]$$

$$\chi^2 = \frac{(ad-bc)^2}{N}\left[\frac{N^2}{(a+b)(a+c)(b+d)(c+d)}\right]$$

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

Hence proved.

YATES:

if any of the cell frequencies C_{ij} <5 Yates has introduced a correction term in the formula for χ^2 2X2 contingency table. It is given by

$$\chi^2 = \frac{N\left[|ab-bc| - N/2\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

Problem 22: The theory predicts the proportion of beans in 4 groups A, B, C, and D should be 9:3:3:1. In an experiment of 1600 beans the frequencies in the four groups are 822, 313, 287, 118. Do these experiment results support the theory.

Solution:

Null hypothesis: There is no significant difference between theoretical frequencies and expected frequencies.

Alternative hypothesis: There is significant difference between theoretical frequencies and expected frequencies.

Test statistic:

$$\chi^2 = \frac{\sum_{i=1}^{n}(O_i - E_i)^2}{E_i} \sim \chi^2_{(n-1)}$$

$O_1 = 882 \qquad O_2 = 313 \qquad O_3 = 287 \qquad O_4 = 114$

$E_1 = 1600 \times \frac{9}{16} = 900 \qquad E_2 = 1600 \times \frac{3}{16} = 300 \qquad E_3 = 1600 \times \frac{3}{16} = 300$

$E_4 = 1600 \times \frac{1}{16} = 300$

$\frac{(O_1 - E_1)^2}{E_1} = \frac{(882 - 900)^2}{900} = 0.36 \qquad \frac{(O_2 - E_2)^2}{E_2} = \frac{(313 - 300)^2}{300} = 0.56$

$\frac{(O_3 - E_3)^2}{E_3} = \frac{(287 - 300)^2}{300} = 0.563 \qquad \frac{(O_4 - E_4)^2}{E_4} = \frac{(118 - 100)^2}{100} = 3.24$

$\chi^2 = 0.36 + 0.563 + 0.563 + 3.24$

$\chi^2_{cal} = 4.726$

$\chi^2_{(4-1)} = \chi^2_{(3),\frac{\alpha}{2}} = 7.815$

Inference: since $\chi^2_{cal} = 4.726 < \chi^2_{(3),\frac{\alpha}{2}} = 7.815$ we accept the null hypothesis. There is no significant difference between theoretical frequencies and expected frequencies.

Problem 23: Find following table gives the number of aircraft accidents that occurred during the seven days of week. Find whether the accidents are uniformly distributed overly week.

Days	sunday	monday	tuesday	wednesday	thursday	friday	Saturday
No.of.accidents	16	14	18	12	11	15	14

Solution:

Null hypothesis: The accidents are uniformly distributed overly the week.

Alternative hypothesis: The accidents are not uniformly distributed overly the week.

Level of significance: $\alpha = 0.05$

Test statistic:

$E_1 = 100 \times \frac{1}{7} = 14.2857$

$E_2 = 100 \times \frac{1}{7} = 14.2857$

$\frac{(O_1 - E_1)^2}{E_1} = \frac{(16 - 14.2857)^2}{14.2857} = 0.2057$

$$\frac{(O_2-E_2)^2}{E_2}=\frac{(14-14.2857)^2}{14.2857}=0.0059$$

$$\frac{(O_3-E_3)^2}{E_3}=0.9657 \qquad \frac{(O_{\backslash 4}-E_4)^2}{E_4}=0.3657$$

$$\frac{(O_5-E_5)^2}{E_5}=0.7557 \qquad \frac{(O_{\backslash 6}-E_6)^2}{E_6}=0.3657$$

$$\frac{(O_7-E_7)^2}{E_7}=0.0057$$

$$\chi^2_{cal}=2.3399$$

$$\chi^2_{\frac{\alpha}{2}(6)}=12.592$$

$$\chi^2_{cal}=2.3399<\chi^2_{\frac{\alpha}{2}(6)}=12.592$$

We accept the null hypothesis and therefore the accidents are uniformly distributed overly the week.

TEST FOR HOMOGENEITY OF SEVERAL CORRELATION COEFFICIENT:

Assumption:

1. The population is bivariate normal with correlation coefficient $\rho_1,\rho_2,\cdots\rho_k$
2. k random samples are drawn and their correlation coefficient are denoted by $r_1,r_2,\cdots r_k$ the samples are large.

Null hypothesis: there is no significant difference between several correlation coefficient

$$H_0:\rho_1=\rho_2=\cdots=\rho_k=\rho$$

Alternative hypothesis: there exist atleast one of the correlation coefficient unequal.

$$H_1:\rho_1\neq\rho_2\neq\cdots=\rho_k\neq\rho$$

Level of significance: $\alpha=0.05$

Test statistic:

$$Z_i=\frac{1}{2}\log\left(\frac{1+r_i}{1-r_i}\right) \quad ;i=1,2,\ldots,k$$

$$Z_i=\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$$

$$Z_i \sim N\left(\varepsilon, \frac{1}{n_i - 3}\right)$$

r_i is the ith sample correlation coefficient.

n_i is the ith sample size

ρ is the population correlation coefficient under H_0.

Since – is unknown it is estimated by

$$\overline{Z} = \frac{\sum_{i=1}^{k} Z_i (n_i - 3)}{\sum_{i=1}^{k} (n_i - 3)}$$

The test statistic is,

$$= \sum_{i=1}^{k} \left(\frac{z_i - \overline{z}}{1/\sqrt{n_i - 3}} \right)^2 \sim \chi^2_{(k-1)}$$

Critical value:

From χ^2 table we can find the value $\chi^2_{(k-1)}$ for (k-1) degrees of freedom.

Inference:

$\chi^2_{cal} > \chi^2_{\alpha,(k-1)}$ We reject H_0 otherwise there is no reason to reject H_0.

UNIT V

In this chapter will discuss the nonparametric procedures for inference.

NON-PARAMETRIC INFERENCE

Most of the standard methods of statistical inference are based on the familiar assumption that the random variables have normal distributions. Then the given procedures are optimum. But for non-normal distributions the standard procedures may be far from optimum. In such cases non-parametric methods are used. A procedure will be called distribution free if the statistic used has a distribution which does not depend on the distribution function of the population from which the sample is drawn. So in such procedures assumptions regarding the population are not necessary.

Distinguish between parametric and non-parametric:

In parametric test we are concerned with testing parameters of the population t, F, χ^2 and normal test are used to find the parametric values.

The features of parametric test are null hypothesis are defined using the values the values of the parameter for e.g

In parametric test we are concerned with testing parameters of the population t, F, χ^2 and normal test are used to find the parametric values.

The features of parametric test are null hypothesis are defined using the values the values of the parameter for e.g $\mu = 20, \sigma = 5, \rho = 0$

The form of population is assumed to be known samples drawn from the population are independent and random.

The sampling distribution of the statistic is wither exactly known or asymptotically calculated.

When the form of population is unknown we cannot apply t, F, χ^2

When test are not based on the form of distribution we have distribution free test or non-parametric test.

When the parameters are not tested we are interested in testing any measure of location or whether the two population have the same density function, in such cases we use non-parametric test.

Non-parametric:

Assumption:

- The form of the population is unknown.
- The population possesses density function.
- Lower order moments exists i.e μ_1', μ_2' are finite.
- Sample observation are independent and random
- The variable under study is continuous.
- The two population are identical or the measures of location of two population are the same.

RUN TEST:

A run test is sequence of letter of one kind followed by a sequence of letters of another kind. The number of letters in a sequence is called length of the run.

For example: xxx/y/xxxxxx/y/x/yyyy

The sequence has 6 runs. The length of third run is 6.

Let $x_1, x_2, \cdots x_n$ and $y_1, y_2, \cdots y_n$ be two random samples from two given population.

Null hypothesis: Are the two populations having identical density function.

$H_0 : f_x(\bullet) = g_y(\bullet)$

Alternative hypothesis: the two populations do not have identical density function.

$H_1 : f_x(\bullet) \neq g_y(\bullet)$

Level of significance: $\alpha = 0.05/0.01/ any\, other\, specified\, value$

Test statistic and its distribution:

Let U denote the number of rums in the combined sample.

$$E(U) = \left(\frac{2n_1 n_2}{n_1 + n_2} \right) + 1$$

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

The above – and – are computed on the basis of large sample.

$$Z_0 = \frac{U - E(U)}{\sqrt{v(U)}} \sim N(0,1)$$

Critical value:

$$\alpha = 0.05, z_{\frac{\alpha}{2}} = 1.96$$

$$\alpha = 0.01, z_{\frac{\alpha}{2}} = 1.96$$

Inference: if $Z_{cal} > Z_{\frac{\alpha}{2}}$ we reject H_0 otherwise there is no reason to reject H_0

Problem 1: The following data relates to two population observations

SI	10	20	15	25	18	28	23	10	12	14	
SII	11	13	18	28	30	32	24	27	22	11	12

Test whether the samples have come from sample population.

Solution:

10 10/ 11 11/12 12/13/14/15/18 18/20/22 22/23/24/25/27/28 28/30/32

Null hypothesis: the samples are come from sample population.

Alternative hypothesis: the sample are not come from sample population.

Level of significance: $\alpha = 0.05$

Test statistic:

$$U = 14 \quad n_1 = 10 \qquad n_2 = 12$$

$$E(U) = \left(\frac{2(10)(12)}{10 + 12} \right) + 1$$

$$E(U) = 11.9091$$

$$V(U) = \frac{2(10)(12)(240 - 10 - 12)}{(10 + 12)(10 + 12 - 1)}$$

$$V(U) = 5.1476$$

$$Z_0 = \frac{14 - 11.9091}{\sqrt{5.1476}}$$

$Z_0 = 0.9216$

$Z_{\alpha/2} = 1.96$

$Z_0 = 0.9216 > Z_{\alpha/2} = 1.96$

We reject the null hypothesis. Therefore the sample are not come from sample population.

TEST FOR RANDOMNESS, RUN TEST:

Let $x_1, x_2, \cdots x_n$ be sample from the given population.

- Find the median of the sample.
- Represent the given observation in the same order by A or B. where A stands for above median and B stands for below median.
- If any observation is equal to median omit the observation and reduce the size of the sample.
- Find the number of runs and denote it by U.
- Expectation of U ie $E(U) = \frac{n+2}{2}$ *and* $V(U) = \frac{n}{4}\left(\frac{n-2}{n-1}\right)$

$$Z_0 = \frac{U - E(U)}{\sqrt{V(U)}} \sim N(0,1)$$

Null hypothesis: the sample is random

Alternative hypothesis: the sample is not random.

Test statistic:

$$Z_0 = \frac{U - E(U)}{\sqrt{V(U)}} \sim N(0,1)$$

Critical value: for $\alpha = 0.05$ $, Z_\alpha = 1.65$ $, \alpha = 0.01$ $, Z_\alpha = 2.33$

Inference: if $Z_0 > Z_\alpha$ reject H_0 otherwise there is no reason to reject H_0

MEDIAN TEST

Procedure:

- Two samples they namely $x_1, x_2, \cdots x_n$ and $y_1, y_2, \cdots y_n$ are given. Combine the samples and arrange them in ascending order of magnitude and find median.
- Find m_1 (i.e) number of values in the first sample exceeding the median.
- Find m_2 (i.e) number of values in the second sample exceeding the median.

- Form a contingency table as follows:

Samples	No.of.observation above median	No.of.observation below median	Total
1.	m_1	$n_1 - m_1$	n_1
2.	m_2	$n_2 - m_2$	n_2
3.	$m_1 + m_2$	$(n_1 + n_2) - (m_1 + m_2)$	$n_1 + n_2$

Null hypothesis: the two have the same median.

Alternative hypothesis: the two samples do not have the same median.

Level of significance: $\alpha = 0.05 / 0.01 / \textit{any other specified value}$

Test statistic:

$$\chi^2_{cal} = \sum_{i=1}^{m} \sum_{j=i}^{n} \left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right) \sim \chi^2_{(m-1)(n-1)}$$

$$\frac{R_i C_j}{N = n_1 + n_2} = E_{ij}$$

m- number of rows n-number of column

Critical value: we can find $\chi^2_{(m-1)(n-1)}$ from χ^2 for (m-1)(n-1) degrees of the freedom at level of significance.

Inference : $\chi^2_{cal} > \chi^2_{\alpha,(m-1)(n-1)}$ we reject null hypothesis otherwise there is no reason to reject null hypothesis

Note: median test can be used for testing equality of medians of any number of population. Suppose there are r samples we get r/2 contingency table. Therefore the degrees of freedom for the problem will be (r-1)(2-1).

Problem 2: 3 random samples are drawn from 3 population gave the following values if whether the population have the same median.

SI	1	2	5	7	8	9	3	2			
SII	2	5	3	8	9	5	2	7	10		
SIII	3	4	2	5	7	8	9	7	11	8	12

Solution:

Null hypothesis: the three samples have the same median

Alternative hypothesis: the three samples do not the same median.

Level of significance: $\alpha = 0.05$

Test statistic:

Median= (5+7)/2=6

Samples	No.of.observation above median	No.of.observation below median	Total
1	3	5	8
2	4	5	9
3	7	4	11
Total	14	14	28

O_{ij}	E_{ij}	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
3	4	0.25
5	4	0.25
4	4.5	0.0556
5	4.5	0.0556
7	5.5	0.4091
4	5.5	0.4091

χ_0^2 =1.4294

$\chi^2_{(2)(1)} = 5.99$

$\chi_0^2 = 1.4294 < \chi^2_{(2)(1)} = 5.99$

We accept null hypothesis, the three samples have same median.

SIGN TEST:

Sign test is preferred under the following situations

- Population density function is unknown
- Sample observations are paired
- Different pairs are observed under different variances and so paired –t test cannot be applied.

- Measurements are such that d_i=x_i-y_i can be expressed as positive or negative sign.
- Variables are continuous and $d_i{}^{'}$s are independent.

Procedure:

Null hypothesis:

Two populations have identical distribution

ie., $f_x(.)=f_y(.)$,P[(X-Y)>0]=1/2, P[(X-Y)<0]=1/2

Alternative Hypothesis:

Two populations have different distribution

$f_x(.) \neq f_y(.), \quad P[(X-Y)<0] \neq 1/2$

Level of significance:

$\alpha = 0.05/0.01/\, any\, other\, specific\, value$

Test Statistic

E(U)=np=n(1/2)=n/2

V(U)=npq=n(1/2)(1/2)=n/4

When the sample is large, we can have normal approximation

$$Z_0 = \frac{U - E(U)}{\sqrt{V(U)}} = \frac{2U - n}{\sqrt{n}} \sim N(0,1)$$

Critical Value:

When $\alpha = 0.05$, $Z_{\alpha/2} = 1.96$ and $\alpha = 0.01$, $Z_{\alpha/2} = 2.58$

Inference:

If $|Z_0| > Z_{\alpha/2}$ we reject H_0 otherwise there is no reason to reject H_0

Note: d_i=x_i-y_i is no sign attached and the pair is omitted from the sample size n and the reduced sample will have n-1 observation.

Problem 3: A random sample of paired observation is given below (10,11), (11,13), (12,10), (13,13), (14,15), (11,14), (12,13), (13,12), (10,8), (10,13), (14,15), (15,17), (15,13), (11,10), (8,9), (9,9), (11,9), (12,14), (13,11), (11, 11). Apply approximately non-parametric test. Test Whether there is any gain in B=X-Y.

Solution:

Null Hypothesis: There is no gain in B=X-Y

Alternative Hypothesis: There is gain in B=X-Y

Level of Significance: $\alpha = 0.05$

Test Statistic

X	Y	Sign	X	Y	Sign	X	Y	Sign
10	11	-	11	14	-	9	9	0
11	13	-	12	13	-	11	9	+
12	10	+	13	12	+	12	14	-
13	13	0	10	8	+	13	11	+
14	15	-	10	8	-	11	11	0

Reduced Sample Size=Total Number signed observations-Non Signed Observations=20-3=17

$$Z_0 = \frac{2U - n}{\sqrt{n}} = \frac{2(7) - 17}{\sqrt{17}} = -0.7276$$

$$Z_{\alpha/2} = 1.96$$

Inference:

If $|Z_0| = 0.7276 < Z_{\alpha/2} = 1.96$. Therefore, there is no reason to reject H_0. There is no gain in B=X-Y.

MANN-WHITENEY U TEST (RANK SUM TEST)

Mann–Whitney U test (also called the Mann Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon Mann–Whitney test) is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample.

Unlike the t-test it does not require the assumption of normal distributions. It is nearly as efficient as the t-test on normal distributions.

A Wilcoxon signed-rank test is a nonparametric test that can be used to determine whether two dependent samples were selected from populations having the same distribution. A Wilcoxon rank sum test is a nonparametric test that can be used to determine whether two independent samples were selected from populations having the same distribution.

PROCEDURE

Null hypothesis

The populations have the same density function i.e., H_0: $f_x(.)=g_y(.)$

Alternative Hypothesis

The populations do not have the same density function. $H_1 : f_x(.) \neq g_y(.)$

Level of significance:

$\alpha = 0.05/0.01/ any\, other\, specific\, value$

Test statistic:

Combine the two sample and assign rank

T= sum of the ranks in second sample

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

where, n_1=size of the first sample

n_2=size of the second sample

Under Asymptotic condition

$$Z_0 = \frac{U - E(U)}{\sqrt{V(U)}} \sim N(0,1)$$

where $E(U) = \frac{n_1 n_2}{2}$, $V(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$

Critical Value:

When $\alpha = 0.05$, $Z_{\alpha/2} = 1.96$ and $\alpha = 0.01$, $Z_{\alpha/2} = 2.58$

Inference:

If $|Z_0| > Z_{\alpha/2}$ we reject H_0 otherwise there is no reason to reject H_0

Problem 4: The following are values obtained from two samples.

x	1	2	3	5	7	9	11	18	
y	4	6	8	10	12	13	14	15	19

Use Mann-whiteney U test to test whether the populations have same density.

Solution:

Null hypothesis: The population has same density function

Alterative Hypothesis: The population does not have same density function

Level of Significance: $\alpha = 0.05$

Test Statistic:

Arrange the ascending order,

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 18,19

T= Sum of second sample values= 4+6+8+10+12+13+14+15+17=99

$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T$ =72+(90/2)-99=18

$E(U)=\frac{n_1 n_2}{2}$ =36, $V(U)=\frac{n_1 n_2(n_1+n_2+1)}{12}$ =108

$$Z_0=\frac{U-E(U)}{\sqrt{V(U)}}=\frac{18-36}{\sqrt{108}}=-1.7321$$

$Z_{\alpha/2}=1.96$

$|Z_0|=1.7321<Z_{\alpha/2}=1.96$.

Inference:

Therefore, there is no reason to reject H_0. The population has some density function.

KRUSKAL-WALLIS TEST

The Kruskal-Wallis test is a nonparametric (distribution free) test, and is used when the assumptions of one-way ANOVA are not met. Both the Kruskal-Wallis test and one-way ANOVA assess for significant differences on a continuous dependent variable by a categorical independent variable (with two or more groups). In the ANOVA, we assume that the dependent variable is normally distributed and there is approximately equal variance on the scores across groups. However, when using the Kruskal-Wallis Test, we do not have to make any of these assumptions. Therefore, the Kruskal-Wallis test can be used for both continuous and ordinal-level dependent variables. However, like most non-parametric tests, the Kruskal-Wallis Test is not as powerful as the ANOVA.

Assumptions

1. We assume that the samples drawn from the population are random.
2. We also assume that the observations are independent of each other.
3. The measurement scale for the dependent variable should be at least ordinal.

Null hypothesis:

Null hypothesis assumes that the samples (groups) are from identical populations.

Alternative hypothesis:

Alternative hypothesis assumes that at least one of the samples (groups) comes from a different population than the others.

Level of significance:

$\alpha = 0.05/0.01/ any\, other\, specific value$

Test Statistic:

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} \right] - 3(n+1)$$

Where, n = sum of sample sizes for all samples,

c = number of samples,

T_j = sum of ranks in the j^{th} sample,

n_j = size of the j^{th} sample.

Inference : $H_{cal} > \chi^2(c-1)df$, we reject null hypothesis otherwise there is no reason to reject null hypothesis

Problem 5: A shoe company wants to know if three groups of workers have different salaries:

Women: 23K, 41K, 54K, 66K, 78K.

Men: 45K, 55K, 60K, 70K, 72K

Minorities: 18K, 30K, 34K, 40K, 44K.

Solution:

Null Hypothesis: There is no significant different between the salary of three groups of workers

Alternative Hypothesis: There is significant different between the salary of three groups of workers

Level of Significance: $\alpha = 0.05$

Sort the data for all groups/samples into ascending order in one combined set.

18K, 23K, 30K, 34K, 40K, 41K, 44K, 45K,54K,55K,60K,66K,70K,72K,78K

Assign ranks to the sorted data points. Give tied values the average rank.

20K- 1, 23K-2, 30K-3, 34K-4, 40K-5, 41K-6, 44K-7, 45K-8, 54K-9, 55K-10, 60K-11, 66K -12, 70K-13, 72K-14, 90K-15

Add up the different ranks for each group/sample.

Women: 23K, 41K, 54K, 66K, 90K = 2 + 6 + 9 + 12 + 15 = 44.

Men: 45K, 55K, 60K, 70K, 72K = 8 + 10 + 11 + 13 + 14 = 56.

Minorities: 20K, 30K, 34K, 40K, 44K = 1 + 3 + 4 + 5 + 7 = 20.

Test statistic

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} \right] - 3(n+1)$$

$$H = \left[\frac{12}{15(15+1)} \left[\frac{44^2}{5} + \frac{56^2}{5} + \frac{20^2}{5} \right] \right] - 3(15+1) = 6.72$$

ind the critical chi-square value. With c-1 degrees of freedom. For 5 – 4 degrees of freedom and an alpha level of .05, the critical chi square value is 9.4877.

The chi-square value is not less than the test statistic, so there is not enough evidence to suggest that the means are unequal.

Hypothesis Tests of the Mean and Median

Nonparametric tests are like a parallel universe to parametric tests.

Parametric tests (means)	**Nonparametric tests (medians)**
1-sample t test	1-sample Sign, 1-sample Wilcoxon
2-sample t test	Mann-Whitney test
One-Way ANOVA	Kruskal-Wallis, Mood’s median test
Factorial DOE with one factor and one blocking variable	Friedman test