Chapter-I

Introduction
CHAPTER I

INTRODUCTION

1.1 Introduction

The outcome of any experiment or the result of any natural phenomenon depends on many unknown factors, which cannot be completely controlled or measured exactly. It is not possible to explain such situations by deterministic mathematical equations. A better way of studying the behaviour of such phenomena, when the outcomes are affected by many uncertain factors is by using probabilistic models. These are the models defined in terms of random variables. For example, suppose that one wants to know the value of $x$, where $x$ may be price of certain commodity, or the contents of a reservoir, or the velocity of wind, or the amount of currency notes in the Reserve Bank of India, or the stock of radio active material etc. Methods of analysis of data based on probabilistic models are called statistical methods. Statistical methods of analysis are intended to aid the interpretation of data subject to appreciable variation.

Virtually all statistical information procedures are based on assumptions concerning probability properties of the observational data used in their application. The accuracy of
the probability information obtained from a statistical procedure depends on how closely the assumed and actual properties agree. Classical methods for measurement data which are based on the assumption of normality for population samples are often relatively insensitive to the assumption, to have broader applicability, and to be easier to use.

Most statistical work is concerned directly with the provision and implementation of methods for study design and for the analysis and interpretation of data. Theory of statistics deals in principle with the general concepts underlying all aspects of such work.

The concern here is likely to be more concentrated on whether models have been reasonably formulated to address the most fruitful questions, on whether the data are subject to unappreciated errors or contamination and, especially, on the subject matter interpretation of the analysis and its relation with other knowledge of the field.

Statistical inference is the branch of statistics which is concerned with using probability concept to deal with uncertainty in decision making. The entire body of classical statistical inference techniques is based on fairly specific assumptions regarding the nature of the underlying population.
distribution; usually its form and some parameter values must be stated. Given the right set of assumptions, certain test statistics can be developed using mathematics, which is frequently elegant and beautiful. The derived distribution theory is qualified by certain prerequisite conditions, and therefore all conclusions reached using these techniques are exactly valid only so long as the assumptions themselves can be sustained. In a real world problem, everything does not come packaged with labels of population of origin. A decision must be made so as to what population properties may judiciously be assumed for the model. If the reasonable assumptions are not such that the traditional techniques are applicable, the classical methods may be used and inference conclusions stated only with the appropriate qualifiers.

We can broadly classify statistical inference as

a) Parametric Inference

b) Non-parametric Inference

**a) Parametric Inference**

In 1908, student (W. S. Gossef) under the assumption of normality, defined a test statistic to test the equality of means when two samples are independent with common unknown variance (classical t-test or two sample t-test). The t-test has
various optimal properties. It is asymptotically distributed having finite fourth moment. But population distribution may be of such a form that this property is not satisfied. Also in practice, one cannot always assume normality. Two sample t-test fails to find out the difference between the two different distributions whose first two moments are the same.

In 1928, Neyman-Pearson proposed a likelihood ratio test under the assumption that the forms of the distributions from which the samples are drawn are well-know (distributions need not necessarily be normal). The L.R. test is equivalent to students t-test has some optimum properties, the experimenter, many a times will not be in a position to know the form of the distributions from which his observations come from.

The main difficulty with the parametric inference is the possible non validity of one or more of the assumptions. Thus, if one rejects certain hypothesis, it may be due to falsehood of the null hypothesis or due to wrong parametric form of the underlying distribution will lead to either wrong level of significance or power. Thus it is desirable to have an alternative set of procedures that are valid under broad assumptions on the underlying populations. Indeed non
Introduction

Parametric methods are of high efficiency relative to classical techniques under the assumption of normality and often of higher efficiency in other situations.

b) Non-parametric Inference

It is fashionable to claim that non-parametric methods were first used when T. Arbuthnot (1710) found that in each year from 1629 to 1710, the number of males christened in London exceeded the number of females. Spearman (1904) proposed a rank correlation coefficient that bears his name, and Karl Pearson (1911) proposed chi-square test. The systematic development took place from 1945 when Wilcoxon proposed his famous Wilcoxon rank sum test. Non-parametric test can be applied in many practical situations since one need not assume that the samples come from a particular distribution. One has a broader boundary than classical inference. Moreover, since most of the non-parametric test statistics are based on counts and ranks of observations, the experimenter with least mathematical background can use the techniques.

Inference in Reliability

A certain question in reliability theory pertains to the modelling of the probability distributions of random variables
representing the life time of units whether human beings, animals, radioactive substance, components and system of components etc. A popular model is the exponential distribution which is useful whenever the 'no aging' phenomenon is evident. Translated into reliability terms it simply means that the probability distribution of the life time of the unit does not change with the knowledge that the unit has already survived for a given time. As against this, many units exhibit positive aging phenomena. The exponential hypothesis is important because of its implications concerning the random mechanism operated in the experiment being considered. Tests of exponentiality are subject to the usual dilemma concerning the random mechanisms operated in the experiment being considered. Tests of exponentiality are subject to the usual dilemma concerning goodness of fit tests. Only when the hypothesis is rejected do we have a significant result. On the other hand when a test rejects an exponential model it justified the use of other complicated models and probabilistic and statistical methods go along with these models. The term 'positive aging' is used to denote the situation where the performance of a unit deteriorates with its age. Classes of life distributions based on the notion of aging
have been introduced in the literature. Some of the classes of life distributions based on aging are Increasing Failure Rate (IFR), Increasing Failure Rate Average (IFRA), New Better than Used (NBU), New better than used of specified age (NBU-t0) and New better than used in the tail (NBU-[t0,∞]). The chain of implication of these notions is given by

$$\text{IFR} \Rightarrow \text{IFRA} \Rightarrow \text{NBU} \Rightarrow \text{NBU} - t_0 \Rightarrow \text{NBU} - [t_0, ∞]$$

In practical situations it has been noticed that lifetime of units possess one or the other aging property. Hence it is of interest to the statisticians to propose test of the null hypothesis of exponentiality against positive aging.

1.2 Preliminaries

1.2a One Sample U-Statistic

Let $X_1, ..., X_n$ denote a random sample from a distribution with cumulative distribution function $F(x)$. A parameter $\gamma$ is said to be estimable of degree $r$ for the family of distribution $\tau$ if $r$ is the smallest size for which there exists an estimator of $\gamma$ that is unbiased for every $F \in \tau$. That is there exists a function $h(.)$ such that

$$E_{F}[h(X_1, ..., X_r)] = \gamma \quad \text{for every } F \in \tau$$
Without loss of generality the sample kernel $h(\cdot)$ can be assumed to be symmetric in its arguments.

That is

$$h(X_1,\ldots,X_r) = h(X_{a_1},\ldots,X_{a_r})$$

for every permutation $(a_1,\ldots,a_r)$ of integers $1,\ldots,r$. For every kernel $h(X_{a_1},\ldots,X_{a_r})$ there always exists one that is symmetric in its arguments.

$$h'(X_1,\ldots,X_r) = \frac{1}{r!} \sum_{\pi \in A} h(X_{a_1},\ldots,X_{a_r})$$

were the summation is over $A = \{\alpha/\alpha \text{ is a permutation of integers } 1,\ldots,r\}$. $h'(\cdot)$ is symmetric and is an unbiased estimator of $r$ for each $F(x) \in \tau$.

U-statistic for the estimable parameter $\gamma$ of degree $r$ created with the symmetric kernel $h'(\cdot)$ is given by

$$U(X_1,\ldots,X_n) = \frac{1}{\binom{n}{r}} \sum_{\beta} h'(X_{\beta_1},\ldots,X_{\beta_r})$$

where $\beta$ is the collection of all subsets of $r$ integers chosen without replacement from integers $(1,\ldots,n)$. The U-statistic thus constructed is the unique Minimum Variance Unbiased Estimator (MVUE) of $\gamma$. 
1.2b Variance of U-statistic

For a symmetric kernel \( h^*(.) \) consider the random variables \( h^*(X_1,...,X_c,X_{c+1},...,X_r) \) and \( h^*(X_1,...,X_c,X_{r+1},...,X_{2r-c}) \) having exactly \( c \) variables in common. The covariance between these two variables is given by

\[
\xi_{2c} = \text{Cov}\left[ h^*(X_1,...,X_c,X_{c+1},...,X_r) h^*(X_1,...,X_c,X_{r+1},...,X_{2r-c}) \right]
= \mathbb{E}\left[ h^*(X_1,...,X_c,X_{c+1},...,X_r) h^*(X_1,...,X_c,X_{r+1},...,X_{2r-c}) \right] - \gamma^2
\]

Further

\[
\xi_{2c} = \text{Cov}\left[ h^*(X_{\beta_1},...,X_{\beta_r},X_{\beta'_1},...,X_{\beta'_r}) \right]
\]

where \( (\beta_1,...,\beta_r)^' \) and \( (\beta'_1,...,\beta'_r)^' \) are subsets of the integers \( \{1,2,...,n\} \) having exactly \( c \) integers out of \( r \) in common.

Then,

\[
\text{Var}(U) = \frac{1}{n} \sum_{c=1}^{r} \binom{r}{c} \binom{n-r}{r-c} \xi_{2c}
\]

Asymptotic Variance of U-statistics

Let \( U(X_1,...,X_n) \) be the U-statistic for the symmetric kernel \( h^*(X_1,...,X_r) \).

If \( \mathbb{E}[h^*(X_1,...,X_r)] < \infty \), then

\[
\lim_{\eta \to \infty} n \text{Var}[U(X_1,...,X_n)] = r^2 \xi_{2}\text{.}
\]
One Sample U-statistic Theorem (Hoeffding, 1948)

Let $U(X_1,...,X_n)$ be a random sample from a distribution $F \in \tau$ and $\gamma = \gamma(F)$ be an estimable parameter of degree $r$ with symmetric kernel $h^*(X_1,...,X_r)$.

Define

$$U(X_1,...,X_n) = \frac{1}{n} \sum_{h \in \beta} h^*(X_{\beta_1},...,X_{\beta_r})$$

where $\beta$ is the collection of all subsets of $r$ integers chosen without replacement from integers $(1,...,n)$.

If $\xi_1 > 0$ then

$$\sqrt{n} (U - \gamma) \xrightarrow{D} N(0, r^2 \xi_1)$$

For application purpose $\xi_1$ is computed using the relation.

$$\xi_1 = \text{Var}[h^*_1(X_1)]$$

where

$$h^*_1(x) = E[h^*(X_1,X_2,...,X_r)]$$

Covariance between Two U-statistics (Lee, 1990)

Let $U_n^{(1)}$ and $U_n^{(2)}$ be two U-statistics, both based on a common sample $X_1,...,X_n$ but having different kernels $\psi$ and $\phi$ of
degrees $k_1$ and $k_2$ respectively, with $k_1 \leq k_2$. Let $\sigma_{c,d}^2$ be the covariance between the conditional expectations.

$$\psi_c(X_1,\ldots,X_c) = E\{\psi(X_1,\ldots,X_c,X_{c+1},\ldots,X_{k_1})\}$$

and

$$\phi_d(X_1,\ldots,X_d) = E\{\phi(X_1,\ldots,X_d,X_{d+1},\ldots,X_{k_1})\}$$

and if $S$ is a set, Let $|S|$ denote the number of elements in $S$.

Theorem 1: Suppose that $c \leq d$. If $S_1$ is in $S_{n,k_1}$, and $S_2$ is in $S_{n,k_2}$ with $|S_1 \cap S_2| = c$, where $S_{n,k_1}$, denotes the set of $k_1$ subsets of $\{1,\ldots,n\}$ and $S_{n,k_2}$, denote the set of $k_2$ subsets of $\{1,\ldots,n\}$ then

$$\sigma_{c,d}^2 = Cov(\psi(S_1),\phi(S_2))$$

Theorem 2: The covariance between $U_n^{(1)}$ and $U_n^{(2)}$ as defined above is given by

$$Cov(U_n^{(1)},U_n^{(2)}) = \binom{n}{k_1}^{-1} \sum_{c=1}^{k_1} \binom{k_2}{c} \binom{n-k_2}{k_1-c} \sigma_{c,c}^2$$

**Asymptotic Covariance between two U-statistics**

Let $X_1,\ldots,X_n$ be a random sample from a cumulative distribution function $F(x)$. Let $U_1$ be a U-statistics for an estimable parameter of degree $c$ with kernel $\psi(.)$ and $U_2$ be
another U-statistic for another estimable parameter of degree 
d with kernel \( \phi(.) \).

Then,

\[
\lim_{{n \to \infty}} n \operatorname{Cov}(U_1, U_2) = c d \sigma^2_{i,1},
\]

where

\[
\sigma^2_{i,1} = \operatorname{Cov}[\psi^*(X_1, \ldots, X_c), \phi^*(X_1, X_{c+1}, \ldots, X_{c+d-1})]
\]

Refer Lee (1990).

1.2c Two-sample U-statistics

Let \( X_1, X_2, \ldots, X_m \) and \( Y_1, Y_2, \ldots, Y_n \) be two independent 
random samples from distributions with cdf’s \( F(x) \) and \( F(y) \) 
respectively. A parameter \( \gamma \) is said to be estimable of degree 
\((r,s)\), for distributions \((F,G)\) in a family \( \Gamma \) if \( r \) and \( s \) are the 
smallest sample sizes for which there exists an estimator of \( \gamma \) 
that is unbiased for every \((F,G) \in \Gamma\), that is there is function 
\( h^*(.) \) such that.

\[
E_{F,G}[h^*(X_1, \ldots, X_r; Y_1, \ldots, Y_s)] = \gamma \text{ for every } (F,G) \in \Gamma,
\]

without loss of generality. The two sample kernel \( h(.) \) can be assumed 
to be symmetric in its \( X_i \) components. Letting \( h(.) \) denote such 
a symmetric two sample kernel, we have the following direct 
extension of the concept of a U-statistic to this two-sample 
setting. For an estimable parameter \( \gamma \) of degree \((r,s)\) and with
symmetric kernel \( h(\cdot) \), a two sample U-statistic has for \( m \geq r \)
and \( n \geq s \), the form

\[
U_{mn} = U(X_1, \ldots, X_m; Y_1, \ldots, Y_n) = \frac{1}{\binom{m}{r} \binom{n}{s}} \sum_{\alpha \in A} \sum_{\beta \in B} h(X_{\alpha_1}, \ldots, X_{\alpha_r}; Y_{\beta_1}, \ldots, Y_{\beta_s})
\]

Where \( A[B] \) is the collection of all subsets of \( r[s] \) integers
chosen without replacement from the integers \( \{1, \ldots, m\} \)
\( [(1, \ldots, n)] \).

For integers \( c \) and \( d \) such, that \( 0 \leq c \leq r \) and \( 0 \leq d \leq s \), let
\( \xi_{c,d} \) denote the covariance between two kernels random
variables with exactly \( c X_i \)'s and \( d Y_j \)'s in common, that is, let

\[
\xi_{c,d} = \text{Cov}[h(X_1, \ldots, X_r, Y_1, \ldots, Y_d, Y_{d+1}, \ldots, Y_s),
\]

\[
h(X_1, \ldots, X_c, Y_{c+1}, \ldots, X_{2r-c}; Y_1, \ldots, Y_d, Y_{s+1}, \ldots, Y_{2s-d})]
\]

Then, the exact expression for the variance of \( U \) is

\[
\text{Var}[U_{mn}] = \frac{1}{\binom{m}{r} \binom{n}{s}} \sum_{c=0}^{r} \sum_{d=0}^{s} \binom{r}{c} \binom{m-r}{r-c} \binom{s}{d} \binom{n-s}{s-d} \xi_{c,d}
\]

with \( \xi_{0,0} = 0 \)

The following generalized U-statistics theorem gives the
asymptotic distribution of \( U \).
Theorem: Let $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ denote independent random samples from populations with cdf's $F(x)$ and $G(y)$ respectively. Let $h(.)$ be a symmetric kernel for an estimable parameter $\gamma$ of degree (r.s). If $E[h^2(X_1,\ldots,X_r;Y_1,\ldots,Y_s)]<\infty$, then $\sqrt{N}[U(X_1,\ldots,X_m;Y_1,\ldots,Y_n)-\gamma]$ has a limiting normal distribution with mean 0 and variance $\frac{r^2\xi_{10}}{\lambda} + \frac{s^2\xi_{01}}{1-\lambda}$ provided this variance is positive, where $0<\lambda = \lim_{N\to\infty} \frac{m}{n} < 1$ and $N = m + n$.

In this thesis we have considered one and two sample testing problems which have applications in reliability studies.

The first problem considered in the thesis is the problem of robustness of sequential probability ratio test when the underlying distribution of the population sampled is either exponential with location and scale parameters or gamma with scale shape parameters. The study is concerned with the problem of testing simple hypothesis against simple alternatives for scale parameter of exponential and gamma distributions assuming known location and shape parameter respectively. Here, we study the robustness of the test procedure when location parameter in exponential distribution has undergone a change. Also we study the robustness of the
test procedure when the shape parameter of the Gamma distribution has undergone a change.

The second problem considered is the problem of testing exponentiality against new better than used class of life distributions. Hollander and Proschan (1972) is the first to consider this testing problem followed by Koul (1978), Kumazawa (1983), Ahmad (1994, 2004) and others. Here we have proposed a new test statistic and studied their distributional properties. The asymptotic performance the proposed test statistic is also studied in terms of Pitman asymptotic relative efficiencies (ARE) in comparison with the existing tests.

Next we have considered the two sample problems of testing the equality of two life distributions against the alternative that one distribution possess 'more positive ageing' property than the other distribution. The particular positive ageing property considered in the thesis are 'more Increasing Failure Rate (IFR)' , 'more New Better than Used (NBU)' and 'more new better than used of specified age $t_0$ (NBU-$t_0$)' . Hollander, Park and Proschan (1986) proposed a test procedure for testing equality of two life distributions against the alternative that one distribution possess 'more NBU'
property than the other distribution. Tiwari and Zalkikar (1988) considered the two sample problem of testing against the alternative that one distribution is 'more Increasing Failure Rate Average (IFRA)' than the other distribution. Recently, the problem of testing against the alternative that one distribution is 'more new better than used at specified' than the other distribution is considered by Lim, Kim and Park (2005). In this thesis, we have proposed test procedures for testing equality of two life distributions against the alternative that

(i) One distribution is 'more IFR' than the other distribution,

(ii) One distribution is 'more NBU' than the other distribution and

(iii) One distribution is 'more NBU' than the other distribution at specified age.

The distributional properties of the proposed tests for various alternatives mentioned above are studied and the asymptotic performance of the test procedures is also studied.

1.3 Chapter Wise Summary

In chapter II we consider the problem of testing simple hypothesis against simple alternatives for scale parameter of exponential and gamma distributions assuming known location and shape parameter respectively. Here, we study the
robustness of the test procedure when location parameter in exponential distribution has undergone a change by through the Operating Characteristic (O.C). Also we study the robustness of the test procedure when the shape parameter of the Gamma distribution has undergone a change through the Operating Characteristic (O.C) and Average Sample Number (A.S.N.) curves.

Two sample problem of testing equality of two life distributions against the alternative that one distribution has 'more IFR' property than the other distribution is considered in Chapter III. A test procedure for this problem is proposed and its performance is studied in terms of Pitman asymptotic efficacy. As to the best of our knowledge, no test procedure for this test is available in the literature.

The problem of testing exponentiality against new better than used alternatives has been the problem of interest to the statisticians. The pioneer work in this area is the test due to Hollander and Proschan (1972) followed by Koul (1977), Kumazawa (1983), Ahmad (1994, 2004), Pandit and Anuradha (2007a,b) and others. We study a test procedure for this problem which is based on sub sample minima in chapter IV. The test proposed for one sample set up is extended to two
sample problem of testing equality of two life distributions against the alternative that one distribution possess 'more NBU' property than the other distribution. The asymptotic distribution of one and two sample test procedures proposed in this chapter is studied. Their asymptotic performance in comparison with Hollander, Park and Proschan (1986) is also studied for various alternative distributions.

In chapter V, we propose another test procedure for the problem considered in chapter IV. The distributional properties of the test statistic are studied and the asymptotic performance of the test procedure in comparison with Hollander, Park and Proschan (1986) is studied.

Chapter VI is devoted to the problem of testing equality of two life distributions against the alternative that one distribution possess 'more NBU' property at specified age than the other distribution. A test statistic for this problem is proposed and the distributional properties are studied. The asymptotic relative efficiencies of the newly proposed test relative to the test due to Lim, Kim and Park (2005) is studied for different alternatives.