1.1 INTRODUCTION:

Statistical inference is the branch of statistics which is concerned with using probability concept to deal with uncertainty in decision-making. We can broadly classify statistical inference as

1) Parametric Inference
2) Nonparametric Inference

PARAMETRIC INFERENCE:

In 1908, Student (W.S. Gosset) 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 distribution-free in the class of distributions having finite fourth moment. But population distributions 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 known (distributions need not necessarily be normal). The L.R. test is equivalent to Student's t-test when the underlying distributions are normal. Though this 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 nonvalidity 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 assumptions. Incorrect assumption about the 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. This enquiry led the statistician to non-parametric and or distribution-free inference.

**NONPARAMETRIC TESTS:**

Nonparametric inference has its origin in early 20th century (1911) with the proposition of chi-square test by Karl Pearson. The systematic development took place from 1945 when
Wilcoxon proposed his famous Wilcoxon rank sum test. Nonparametric 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 nonparametric test statistics are based on counts and ranks of observations, the experimenter with least mathematical background can use the techniques.

Since the field is quite wide, we have restricted ourselves to studying few problems in this thesis. We have confined ourselves mainly to the study of two-sample location and two sample and k-sample scale problems and multiple comparisons of treatments with controls. In each chapter we have proposed a new test statistic and evaluated its performance. The performances of these statistics have been studied in terms of Pitman asymptotic relative efficiency and consistency.

1.2 PRELIMINARIES:

i) TWO -SAMPLE LOCATION PROBLEM :

The two-sample location problem is one of the fundamental problems encountered in statistics. This problem arises when one would like to know whether two samples come from the same distribution or they come from distributions which differ
only in location. The problem is encountered in many fields like botany, zoology, medicine, psychology, economics etc.

**DEFINITION OF THE PROBLEM:**

Suppose $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ are independent random samples from absolutely continuous distribution functions $F(x)$ and $G(y)$ respectively where $G(y) = F(y-\theta)$. Then the parameter $\theta$ is known as location parameter. When $\theta > 0$, $Y$'s are stochastically larger than $X$'s, that is, the $Y$ distribution is shifted to the right. If $\theta < 0$, then $X$'s are stochastically larger than $Y$'s, that is, the $X$ distribution is shifted to the right. One wishes to test the hypothesis.

$$H_0 : \theta = 0 \text{ Vs. } H_1 : \theta > 0 \text{ or } \theta < 0 \text{ or } \theta \neq 0$$

which implies testing

$$H_0^* : F(x) = G(x) \text{ Vs. } H_1^* : F(x) \geq G(x) \text{ or } F(x) \leq G(x) \text{ or } F(x) \neq G(x)$$

with strict inequality for atleast one $x$. The first two are one-sided hypotheses and the last one is referred to as two-sided hypothesis.

**ii) TWO -SAMPLE SCALE PROBLEM:**

Another fundamental problem encountered in statistics is that of testing the equality of two population parameters that
measure variability. The equivalent usages for variability are dispersion, spread, scatter and scale.

Suppose $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are independent random samples from continuous populations with cdfs $F(x - \theta_1)$ and $F(y - \theta_2)$ respectively, where $\theta_1, \theta_2$ denote the median of the distributions $X_i, Y_j$ and $\sigma > 0$. Then when $\theta_1$ and $\theta_2$ are known, we wish to test $H_0: \sigma = 1$ Vs. $H_1: \sigma > 1$ (or $\sigma < 1$, or $\sigma \neq 1$), that is, we are testing to see the $Y_j$'s tend to be more spread out than the $X_i$'s. This problem is referred to as a two-sample scale problem. The parameter $\sigma$ is referred to as a scale parameter. We shall assume that $\theta_1 = \theta_2$.

In the normal model, that is when $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are independent and normally distributed with means $\mu_1$ and $\mu_2$ and unknown variances $\sigma_1^2$ and $\sigma_2^2$ respectively with

\[
\text{Var}(X_i) = \sigma_1^2
\]
\[
\text{Var}(Y_j) = \sigma_2^2
\]

the hypothesis $\sigma_1^2 = \sigma_2^2$ is tested using classical F-test defined by

$F = \frac{S_x^2}{S_y^2}$

where

$S_x^2 = \frac{\sum_{i=1}^{m} (X_i - \bar{X})^2}{(m-1)}$,

$S_y^2 = \frac{\sum_{j=1}^{n} (Y_j - \bar{Y})^2}{(n-1)}$, \[5\]
\[
\overline{X} = \frac{\sum_{i=1}^{m} X_i}{m}
\]
and
\[
\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}.
\]

In the nonparametric setup, we can test \(H_0\) vs. \(H_1\) with a two-sample linear rank statistic

\[
S_N = \sum_{j=1}^{N} a_N(R_j)
\]

where \(a_N(i)\) is nonincreasing (nondecreasing) for \(i \leq \frac{N+1}{2}\) and nondecreasing (nonincreasing) for \(i \geq \frac{N+1}{2}\). The following are few notable members of this class. Mood's (1954) statistic \(M\), uses scores defined by

\[
a_N(i) = \left[ i - \frac{(N+1)}{2} \right]^2.
\]

This test gives more weight to extreme ranks.

Siegal – Tukey proposed a statistic (ST) with scores

\[
a_{ST}(i) = \begin{cases} 
2i & \text{for } i \text{ even, } 1 < i \leq N/2 \\
2i - 1 & \text{for } i \text{ odd, } 1 < i \leq N/2 \\
2(N-i)+2 & \text{for } i \text{ even, } N/2 < i \leq N \\
2(N-i)+1 & \text{for } i \text{ odd, } N/2 < i \leq N
\end{cases}
\]

Capon (1961) defined normal scores statistic (NS) with scores

\[
a_{NS} = E \left[ Z_{(i)}^2 \right]
\]
where $Z^{(i)}$ is the $i^{th}$ order statistic in a sample of size $N$ from a standard normal distribution. This can be used when the extreme rankings give dispersion information than the central rankings, that is, when distributions have light tails.

Klotz (1962) defined a quartile normal scores test $K$ with the scores

$$a_k(i) = \left\{ \Phi^{-1}\left( \frac{i}{N+1} \right) \right\}^2$$

where $\Phi$ is cdf for standard normal distribution.

This test gives more weight than that of $M$ test to the extreme ranks.

All these tests listed above require the assumption that the medians of the distributions are same that is $F(0) = \frac{1}{2}$. These are not appropriate for distributions with all mass on the positive axis, that is $F(0) = 0$.

Deshpande and Kalpana Kusum (1984) proposed a test statistic under the assumption that the two distribution functions have a common quartile $a$ ($a$ not necessarily equal to $\frac{1}{2}$). The statistic proposed by them is based on $U -$ statistic

$$T_1 = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j)$$
where

\[
\varphi \left( X_i, Y_j \right) = \begin{cases} 
1 & \text{if } 0 \leq X_i \leq Y_j \quad \text{or} \quad Y_j \leq X_i \leq 0 \\
0 & \text{if } X_i \leq 0 \leq Y_j \quad \text{or} \quad Y_j \leq 0 \leq X_i \\
-1 & \text{if } 0 \leq Y_j \leq X_i \quad \text{or} \quad X_i \leq Y_j \leq 0
\end{cases}
\]

Kalpana Kusum (1985) modified Deshpande – Kusum test statistic. The proposed test statistic is

\[
T_2 = \frac{1}{\binom{m}{2} \binom{n}{2}} \sum_{i_1 \leq i_2} \sum_{j_1 \leq j_2} \sum_{k \leq \frac{m+n}{2}} \varphi \left( X_{i_1}, X_{i_2}; Y_{j_1}, Y_{j_2} \right)
\]

where

\[
\varphi \left( x_1, x_2; y_1, y_2 \right) = \begin{cases} 
1 & \text{if } 0 \leq \max(x_1, x_2) \leq \max(y_1, y_2) \text{ and } x_1, x_2, y_1, y_2 \geq 0 \\
0 & \text{if } 0 \leq \max(y_1, y_2) \leq \min(x_1, x_2) \text{ and } y_1, y_2, x_1, x_2 < 0 \\
-1 & \text{if } 0 \leq \min(x_1, x_2) \leq \min(y_1, y_2) \text{ and } x_1, x_2, y_1, y_2 < 0 \\
\text{otherwise}
\end{cases}
\]

Shetty and Pandit (2004) proposed a class of distribution-free tests based on sub sample medians which takes care of outliers at the extremes of both samples. The proposed test statistic is

\[
U_c = \frac{1}{\binom{m}{c} \binom{n}{c}} \sum_{\lambda} h(x_1, x_2, \ldots, x_c; y_1, y_2, \ldots, y_c)
\]

where \( \sum_{\lambda} \) denotes the sum over all \( \binom{m}{c} \binom{n}{c} \) combinations of X and Y sample observations and

\[
h(x_1, \ldots, x_c; y_1, \ldots, y_c) = h^+(x_1, \ldots, x_c; y_1, \ldots, y_c) - h^-(x_1, \ldots, x_c; y_1, \ldots, y_c).
\]
Here

\[ h^+(x_1,\ldots,x_c;y_1,\ldots,y_c) = \begin{cases} 1 & \text{if } 0 < M_x < M_y, \ x_i, y_i > 0; \ i = 1,2,\ldots,c \\ 0 & \text{Otherwise} \end{cases} \]

and

\[ h^-(x_1,\ldots,x_c;y_1,\ldots,y_c) = \begin{cases} 1 & \text{if } 0 < M_y < M_x, \ x_i, y_i < 0; \ i = 1,2,\ldots,c \\ 0 & \text{Otherwise} \end{cases} \]

where \( M_x = \text{median}(x_1,x_2,\ldots,x_c), \ M_y = \text{median}(y_1,y_2,\ldots,y_c) \) and \( c \) denotes the odd positive integer less than or equal to \( \min(m,n) \).

**iii) 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 cdfs \( F(x) \) and \( G(y) \) respectively. A parameter \( \gamma \) is said to be estimable of degree \((r,s)\), for distributions \((F,G)\) in a family \( \mathcal{F} \) if \( r \) and \( s \) are the smallest sample sizes for which there exists an estimator of \( \gamma \) that is unbiased for every \((F,G)\in \mathcal{F} \), that is, there is a function \( h^\ast() \) such that

\[ E_{F,G}[h^\ast(X_1,\ldots,X_r; Y_1,\ldots,Y_s)] = \gamma \]

for every \((F,G)\in \mathcal{F} \). Without loss of generality the two-sample kernel \( h() \) can be assumed to be symmetric in its \( X_i \) components and separately symmetric in its \( Y_j \) 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 \( h() \), a two sample U-statistic has for \( m \geq r \) and \( n \geq s \), the form

\[
U_{mn} = \frac{1}{\binom{m}{r} \binom{n}{s}} \sum_{\alpha \in A} \sum_{\beta \in B} h\left( X_{\alpha}, \ldots, X_{\alpha}, Y_{\beta}, \ldots, Y_{\beta} \right)
\]

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

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 kernel random variables with exactly \( c \) \( X_i \)'s and \( d \) \( Y_j \)'s in common, that is, let

\[
\xi_{c,d} = \text{Cov}\left[ h(X_{1}, \ldots, X_{c}, X_{r+1}, \ldots, X_{r}, Y_{1}, \ldots, Y_{d}, Y_{s+1}, \ldots, Y_{s}) \right],
\]

then the exact expression for the variance of \( U \) is

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

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

The following generalized U-statistics theorem gives the asymptotic distribution of \( U \).
THEOREM: Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ denote independent random samples from populations with cdf's $F(x)$ and $G(y)$ respectively. Let $h(\cdot)$ 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 $N = m + n$ and $\xi_{10}$ and $\xi_{01}$ are as defined in (1.2.1).

1.3 CHAPTER WISE SUMMARY:

In chapter II we consider three classes of distribution-free tests for the special type of two-sample location problem. This type of problem is quite commonly encountered while comparing the performance of two measuring devices. We compare the performances of three classes of distribution free tests for this problem in terms of Pitman asymptotic relative efficiency (ARE). Linear combination of the two classes of proposed tests are considered to determine the optimum weights, which maximize the efficacy of the linear combination. The performance of the linear combination is studied in terms of Pitman ARE. We also prove the consistency property of the proposed test. The paper
based on this chapter has been accepted for publication in Far East Journal of Theoretical Statistics and will appear in volume 17, No. 1, pp. 79-95 of the journal.

Two-sample scale problem is yet another important problem in statistics. Tests due to Deshpande and Kusum (1984) and Kusum (1985) assume that the population $\alpha^{th}$ quantile are equal. However, Deshpande and Kusum (1984) and Kusum (1985) tests reduce to the assumption of equal medians when $\alpha=1/2$. In chapter III we consider this problem and propose a class of tests based on subsample extremes of both the samples. The test statistic $SM(k)$ performs as well and better than other tests in terms of Pitman asymptotic relative efficiency for some standard distributions. The paper based on this chapter has been presented in the International Conference on Interdisciplinary Mathematical and Statistical Techniques held at Lucknow from December 27-29, 2004. It has also been submitted for publication.

A question that the researchers frequently encounter has to do with the equality of k-population parameters that measure dispersion. Such problems occur in medicine, for example, in the study of severe aortic valvular disease requiring prosthetic
valve replacement, in the study of effect of propranolol on severity of myocardial nurosis etc. In such studies the experimenter is often interested in testing equality of scale parameters against ordered scale alternatives. In chapter IV, we consider the weighted linear combination of consecutive two-sample U-statistic defined in chapter III for testing for ordered scale alternatives. The optimal weights are also determined. The asymptotic relative efficiency of the test statistic (in Pitman sense) is also discussed in this chapter.

A common problem in applied research is the comparison of treatment with a control or standard. Such a situation may arise, for example, when an agronomist tests the effect on crop yield of the addition of chemicals to the soil or pharmacologist assays drug samples to determine their potencies. In designing an experiment to measure the effect of such treatments, it is often desirable to include in the experiment a control in the form of either a dummy treatment to measure the magnitude of experimental response in the absence of treatments under investigation or some recognized standard treatments. Solorzano and Spurrier (2001) developed distribution free procedure for simultaneously comparing $k_1$ treatments with $k_2$ controls. In chapter V we have considered a distribution-free
procedure for simultaneously comparing $k_1$ treatment medians to $k_2$ control medians. The distribution-free procedure is based on the two-sample distribution free test statistics proposed by Shetty and Govindarajulu (1988). The marginal and joint distributions of the test statistics have been obtained and this can be used for simultaneous inference problems for one sided alternatives when $k_1 = k_2 = 2$. 