Chapter 1

Introduction and Summary

1.1 Introduction

A statistical model is a probability distribution considered to enable inferences to be drawn and/or decisions to be made on the basis of data. Selecting a statistical model from amongst a class of models is not an easy task, hence to describe a real data set, it is desirable to consider a wide class of distributions. Further, due to advancements in computing facilities, it has become possible to analyze the data by considering large enough/complicated models. Hence, in the recent times analysis of data under the nonparametric model or semi-parametric models has become popular and is widely accepted by the practitioners.

Many a times apart from the observations from an unknown distribution of interest $F(.)$, some additional information (data) about $F(.)$ may be available from a distribution $G(.)$-a functional of $F(.)$. One can utilize this additional information for estimating $F(.)$ and improve the performance of the estimator. For example, the use of censored observations (as additional information) together with complete observations is well documented. Such additional information may be available in many other forms/situations. Sometimes one may have additional information about the form of $F(.)$. For example, $F(.)$ being unimodal symmetric, in such situations
it is desirable that an estimator of it should also have such a property.

In statistical quality control (SQC), monitoring quality of a production process by using charting statistics are essentiality applications of the theory of testing of the statistical hypothesis. As such, a parametric charting procedure will be better than that of a nonparametric charting procedure, provided the assumed parametric model is an appropriate to describe the random behavior of the quality characteristic. Hence, before implementing the existing parametric procedures, which are known to be optimal under the specified assumptions, it is essential to examine the validity of the model for the data sets being evolved. If there is no substantial evidence to ensure the validity of the assumptions required for a parametric model, it is better to use less restrictive nonparametric or semi-nonparametric procedures, with little sacrifice of optimality, which is guaranteed when assumptions for parametric procedure are really valid.

Quite often, in both parametric and nonparametric set-ups, it is assumed that the variables of interest are symmetrically distributed and subsequent results are developed. In the analysis of variance, errors are assumed to be symmetric about zero. In nonparametric testing for location (i.e. Median), the distribution is assumed to be symmetric. While designing the control charts for location, distribution of process characteristic(s) is assumed to be symmetric about a specified target value. One may consider the lack of symmetry itself as an indicative for the process being out of control. In such situations, to identify the out of control state of the process, it is quite necessary (though not sufficient) to develop charting statistics to detect the lack of symmetry of the process characteristic of interest. Hence, study of nonparametric inference related to symmetric models has a lot of potentiality.

1.2 Motivation for the research problem

In the context of the above comments, we have been much motivated by the following concepts or properties which have been used in statistical inference, modelling and
applications.

(i) The concept of symmetry.

(ii) Use of additional information, may be in the form of data or about the form of the distribution itself.

(iii) Distribution-free nature of the rank vector.

(iv) Generation of new classes of densities by introducing additional parameter(s).

In the following section, we review some literature in the context of the concepts or properties indicated above.

1.3 An overview of the relevant literature

Good amount of literature is available on testing symmetry of a univariate distribution. To test symmetry of a distribution about known point $\theta$, without loss of generality $\theta = 0$, a common test is a weighted sign test of the form $\sum_{k=1}^{n} a_k \text{Sign}(X_k)$ and it is studied by Hajek and Sidak (1967). Choice of $a_k$ for Wilcoxon’s signed-rank statistic is rank of $|X_k|$ in a combined sample of absolute values $|X_1|, |X_2|, \ldots, |X_n|$. When $F(.)$ is symmetric about $\theta$, we know that $F(\theta + x) + F(\theta - x) = 1$, for all $x$. To test the hypothesis that $F(.)$ is symmetric about $0$, Butler (1969) has considered a functional $Q_n(x) = n[F_n(x) + F_n(-x) - 1]$, which is a measure of divergence of the estimators of $F(x)$ and $1 - F(-x)$. He has proposed a statistic $B_n = \sqrt{n} \text{Sup}\{|Q_n(x)|; x \leq 0\}$ to test symmetry of $F(.)$ about $0$. He has obtained exact and asymptotic null distribution of $B_n$. However, if a point of symmetry is unknown then $B_n$ is not distribution-free. Hence, one cannot compute critical values and/or p-value. In such situation, bootstrap procedure is implemented by Schuster and Barker (1987) and Arcones and Gine (1991). Feuerverger and Mureika (1977) have used the imaginary part of empirical characteristic function for testing symmetry about the origin. McWilliams (1990) has proposed a distribution-free
test based on a runs statistic for symmetry about known point and shown through simulation that it is more powerful than some competitors. Ahmad and Li (1997) proposed consistent nonparametric tests for testing symmetry about a known median. This test is based on the $L_2$-norm and the kernel method of density estimation. Cheng and Balakrishnan (2004) obtained a modified sign test for symmetry about zero. The test is based on the number of positive observations in right extreme of $k$ observations, when observations are ordered according to their absolute values by retaining their signs. Randles et al. (1980) have proposed a test for testing symmetry of distribution about an unknown point based on number of right and left triples in sub-samples of size 3 from a random sample of size $n$. An ordered triple $X_1 < X_2 < X_3$ of three arbitrary observations is said to be a right triple (left triple) if $X_3 - X_2 > X_2 - X_1$ ($X_3 - X_2 < X_2 - X_1$). A procedure based on sample spacings is proposed by Ekstorm and Jammalamadaka (2007) for testing symmetry of distribution about unknown point. Based on the necessary (but not sufficient) condition for a density to be symmetric, Patil et al. (2011(a)) provided a measure on a scale of $-1$ to $1$ to quantify the amount of asymmetry of a continuous probability density function, with the value zero referring to a symmetric density and $\pm 1$ referring to respectively positively and negatively most asymmetric densities. Patil et al. (2011(b)) have provided a measure of asymmetry based on the necessary and sufficient condition for symmetry.

Also, testing multivariate symmetry (Liu et al. (1999)) of underlying probability density is attempted by many researchers by applying different statistical techniques. It is well known that $X$ is symmetric about 0 if and only if for every vector $l \in \mathbb{R}^p$, $l'X$ is symmetric about 0 (Zhu et al. (1995)). Fang et al. (1998) have developed a projection number theoretic type statistic to test elliptical symmetry. In developing such type of statistic a key point they used is number theoretic method to choose projection directions and then compute statistic using this projected data. Monte Carlo test for multivariate symmetry is proposed by Diks and Tong (1999). Characterization of diagonal symmetry when location is unknown is considered by Szekely
and Sen (2002). Liang et al. (2008) have developed tests for spherical symmetry based on both univariate and multivariate uniform statistics. Ley and Paindaveine (2012) have extended run statistic proposed by McWilliams (1990) for multivariate central symmetry by using data depth-based run statistic.

Sometimes, additional properties of probability density function are known, like being symmetric, convex or concave on its support etc. So, one may use these properties to define the family of distributions and need to confine to the class of estimators satisfying such properties. As such, \( \mathcal{F}^P \), the family of distributions with the additional property \( P \). Obviously, \( \mathcal{F}^P \subseteq \mathcal{F} \). If \( P \) corresponds to symmetry then a symmetric estimator can be obtained by symmetrizing the empirical distribution about a suitable point. Let a random variable \( X \) have a distribution \( F(.) \). Further, suppose \( F(.) \) has an additional property that distribution \( F(.) \) is symmetric about a known point \( \theta \), for \( \theta \in \mathbb{R} \). That is,

\[
F(x) = 1 - F((2\theta - x) -)
\]

for all \( x \in \mathbb{R} \). To estimate the distribution \( F(.) \), let \( X_1, X_2, \ldots, X_n \) be a random sample of size \( n \) from an unknown \( F(.) \), which is symmetric about a known point \( \theta \) (without loss of generality, \( \theta = 0 \)) . The empirical distribution function \( \hat{F}(y) = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i \leq y]} \) is a natural estimator of \( F(y) \) and it is a nonparametric maximum likelihood estimate (NPMLE) of \( F(y) \) (Efron and Tibshirani (1993)). When the distribution is known to be symmetric about zero, Modarres (2002) obtained its NPMLE (\( = \hat{F}^{mle}(y) \), say) and further he has shown that this NPMLE coincides with the symmetrized estimator \( \hat{F}^S(y) = \frac{1}{2} \{ \hat{F}(y) + 1 - \hat{F}(-y) \} \) (Shuster (1975)). Further, it is shown that, though \( \hat{F}(y) \) and \( \hat{F}^{mle}(y) \) are both unbiased for \( F(y) \), relative efficiency of \( \hat{F}^{mle}(y) \) over \( \hat{F}(y) \) is given by

\[
\rho(\hat{F}(y), \hat{F}^{mle}(y)) = \frac{\text{Var}(\hat{F}(y))}{\text{Var}(\hat{F}^{mle}(y))} = \frac{F(-|y|)(1-F(-|y|))/n}{F(-|y|)(1-2F(-|y|))/2n} = \frac{2(1-F(-|y|))}{1-2F(-|y|)} \geq 2.
\]

That is, under symmetry \( \hat{F}^{mle}(y) \) is more efficient than \( \hat{F}(y) \).

If the point of symmetry \( \theta \) is unknown, Schuster (1991) has shown that Galton estimator of \( \theta \) minimizes \( L_1 \)-distance between \( \hat{F}(\cdot) \) and \( \hat{F}(\cdot; \theta) \), where \( \hat{F}(y; \theta) = \frac{1}{n} \sum_{i=1}^{n} I_{[2\theta - X_i \leq y]} \). Therefore, when point of symmetry is unknown, the closest es-
timator of $F(.)$ closest under $L_1$-norm is a symmetrized estimator, symmetrized around Galton estimator. Also, it is shown that Hodges-Lehmann estimator of $\theta$ minimizes $L_2$-distance between $\hat{F}(.)$ and $\tilde{F}(.; \theta)$. Therefore, in this case the closest estimator of $F(.)$ closest under $L_2$-norm is a symmetrized estimator, symmetrized around Hodges-Lehmann estimator. One may also refer to Lo (1985).

When the information gathered is of the form $(X, Y)$, where $X$ is a random sample from distribution $F(.)$ while $Y$ is some additional information about $F(.)$, $Y$ can be utilized along with $X$ in the course of making inference related to $F(.)$. This additional information could be through a related distribution $G(.)$, which is a functional $H(.)$ of $F(.)$ or derived using $F(.)$. Vardi (1982) has considered $G(.)$, the length biased distribution of $F(.)$ given by $G(x) = \frac{1}{\mu} \int_0^x y \, dF(y), x \geq 0$, where $\mu = \int_0^\infty y \, dF(y)$. We are motivated by the estimation method of Vardi (1982), wherein additional information/data is through another random sample of size $n$ from $G(.)$, a length biased distribution of $F(.)$. The nonparametric maximum likelihood estimate (NPMLE) of a lifetime distribution $F(.)$ is obtained on the basis of two independent samples, one of size $m$ from $F(.)$ and the second of size $n$ from $G(.)$, a length biased distribution of $F(.)$.

Also, the estimation of multivariate distributions has become increasingly important in economics and econometrics. A number of methods have been used to estimate multivariate densities. The parametric approach assumes a fully specified class of distributions up to a finite set of parameters; for example, the multivariate normal distribution is commonly used. Although parametric methods are efficient, they perform very poor, if model is not correct. An alternative approach is to estimate the density by using nonparametric estimators. Popular choices of multivariate density estimation include the kernel and the exponential series estimator. These nonparametric estimators seek a functional approximation to the distributions. In this context, some references are: Puri and Sen (1971), Kooperberg (1998), Wu (2009), Bandyopadhyay et al. (2010).

Kernel density estimators are most popular nonparametric estimators of the
probability density function. Silverman (1986) is a classical book on kernel density estimation. For estimation of univariate density on a restricted support and properties of estimators one may refer Wand et al. (1991), Marron and Ruppert (1994), Bagai and Prakash Rao (1995), Guillamon et al. (1999), Bouzemarni and Scaillet (2005), Karunamuni and Alberts (2005a, 2005b and 2006) and Chaubey et al. (2012). If a multivariate density is \( \nu \)-spherical density, \( f(x; \nu(\cdot), g(\cdot)) = g(\nu(x)) \) (Fernandez et al. (1995)), then for given \( \nu(\cdot) \) the shape function \( g(\cdot) \) can be estimated on its restricted support \( \mathbb{R}^+ \) and hence gives an estimator of \( \nu \)-spherical density \( g(\nu(x)) \).

In the literature, there are many methods to generate model based on well known models. Fernandez et al. (1995) proposed \( \nu \)-spherical densities for modeling data. Azzalini and Dalla Valle (1996) defined a multivariate density function
\[
2N_p(x; \mu, \Sigma)\Phi(\alpha'(x - \mu)), x \in \mathbb{R}^p
\]
where \( N_p(x; \mu, \Sigma) \) is a \( p \)-variate normal density function, with mean vector \( \mu \) and covariance matrix \( \Sigma \); \( \Phi(\cdot) \) is a cumulative distribution function of standard normal random variable (r.v.) and \( \alpha \in \mathbb{R}^p \) is a shape parameter. Mudholkar and Hutson (2000) introduced The epsilon-skew normal distribution for analyzing near normal data. Lai (2004) presents an overview of various methodologies to construct various continuous bivariate distributions. Rattihalli and Basugade (2008) have generated a class of multivariate densities by using contour transformation. Rattihalli and Basugade (2009) have introduced the concept of basic contour transform leading to a new decreasing density on \( \mathbb{R}^+ \). Some more references on generation of new class of densities, their characteristics and applications are: Arnold and Lin (2004), Arnold et al. (2008), Azzalini and Capitanio (1999, 2003).

There is a good amount of literature available on statistical quality control to monitor a process quality characteristic, which is necessarily a repetitive application of testing of the statistical hypothesis. In this context, some references are: Montgomery (2003), Boon and Chakraborty (2012), Ghute and Shirke (2012).
1.4 Summary and organization of the research work

The main objectives of this research work undertaken are as follows.

(i) Developing new nonparametric estimators in presence of additional information.

(ii) To study nonparametric estimators when the distribution is known to be symmetric.

(iii) Developing distribution-free tests for testing symmetry of the distribution function.

(iv) Generation of new classes of densities by introducing additional parameter(s).

(v) Developing estimators for asymmetric densities generated in (iv).

(vi) Development and performance study of charting statistics for monitoring symmetry of distribution of process characteristic about a specified point.

The present thesis deals with nonparametric inference related to symmetric models. The thesis is organized into 6 chapters. The contents of chapters are as below.

This chapter gives an introduction to the topic of study as well as a brief review of relevant literature together with summary of the present work undertaken.

Many a times, apart from the observations from an unknown distribution $F(.)$, the distribution of interest, some additional information about $F(.)$, may be available. Such type of additional information may be available in many situations. For example, a manufacturer is interested to assess the quality of the units produced, based on the lifetimes. For this purpose, he may conduct an experiment on $m$ units and obtain lifetimes $X_1, X_2, \ldots, X_m$. Suppose these units are used in a system as a parallel subsystem of $\alpha$ components and the service station has maintained records $Y_1, Y_2, \ldots, Y_n$ on the lifetimes of the $n$ parallel subsystems. Thus, to estimate the
distribution function of lifetime of components, the records $Y_1, Y_2, \ldots, Y_n$ having distribution $G(t) = [F(t)]^\alpha, t \geq 0, \alpha > 0$ provide an additional information. Here, $G(.)$ is a functional of $F(.)$. Patil et al. (2011) discussed the problem of estimation of $F(.)$ when there are $n$ additional observations from $G(t) = [F(t)]^2, t \geq 0$, when $F(.)$ is not necessarily continuous. Rattihalli and Patil (2011) estimated $F(.)$, where additional information is through $G(t) = [F(t)]^\alpha, t \geq 0, \alpha > 0$. The case that $\alpha$ is known and $\alpha$ is unknown are considered separately. $\alpha$ is a nuisance parameter, when it is unknown.

In Chapter-2, nonparametric estimation in the presence of additional supplementary data is considered and the details are reported in Section 2.2. Nonparametric estimation of density function when it has additional properties is studied in Section 2.3. The problem of testing univariate symmetry is considered in Section 2.4. Distribution-free test for testing symmetry is proposed and its performance, as compared to some existing tests, has been carried out by using extensive Monte Carlo simulation. Performance of the proposed test for a typical class of asymmetric distributions with zero median is better than its competitors. Research article based on this work is submitted for its possible publication in Journal of Indian Statistical Association (Rattihalli and Patil (2012)). Concluding remarks are given in Section 2.5.

We know that, statistical model is a probability distribution considered to enable inferences to be drawn and/or decisions to be made on the basis of data. Many standard statistical models exist in the literature which have been in use to fit the data and draw inferences related to entities of interest. However, existing models may be symmetric or not, may not give a very satisfactory fit for a given data set. This may be due to the presence of a typical unaccounted factor present in the data. This demands the need of different methods/techniques for generation of new statistical models by using the existing standard models.

Chapter-3 deals with generation of new statistical models. We have proposed the class of Generalized $v$-spherical densities, by introducing an additional param-
eter into $\nu$-spherical densities of Fernandez et al. (1995). The generalized class is proposed in Section 3.2. Illustrations are given in the Section 3.3 and amongst others, we obtain generalized p-variate normal, t and Laplace densities. In the Section 3.4, we state some properties of generalized $\nu$-spherical densities and develop a method to generate random sample from these distributions. Adequacy of generalized $\nu$-spherical densities over $\nu$-spherical densities is shown by using real data sets in Section 3.5. Research article based on this work is published in Communications in Statistics- Theory and Methods (Rattihalli and Patil (2010)).

Chapter-4 is devoted to address the problem of estimation of the shape function and asymmetry parameter in the class of generalized $\nu$-spherical densities. In Section 4.2 the problems of estimation of the shape function on $(0, \infty)$ without any restrictions is considered and in Section 4.3 its estimation by imposing the non-increasing restriction is considered. An algorithm, viz. Pool Maximum Violation Region Algorithm (PMVRA), is developed to obtain an estimator of shape function with non-increasing restriction. Detail description of the PMVRA is given in Section 4.4. In Section 4.5, we consider estimation of asymmetry parameter of generalized $\nu$-spherical density, where shape function is not necessarily required to be known. To evaluate the performance of PMVRA and Pool Adjacent Violation Algorithm (PAVA) (Robertson, Wright and Dykstra (1988)), extensive simulation study is carried out in Section 4.6. In this section, performance of estimator of asymmetry parameter is also studied. Concluding remarks are given in Section 4.7.

In Chapter-5, we have developed tests for diagonal symmetry by applying three different techniques. In Sections 5.2, 5.3 and 5.4 we have developed test for diagonal symmetry about a specified point by using i) distances and the directions of the observations from an arbitrary plane passing through the origin, ii) the likelihood ratio criterion and iii) the counts of the observations in the $2^k$ regions formed by k planes passing through the origin, respectively. Performance of these tests through extensive simulation is studied in Section 5.5 by considering some class of alternatives and concluding remarks are given in Section 5.6.
In the context of statistical quality control (SQC), suppose that a process is said to be under statistical control, if its quality characteristic has a symmetric distribution about a specified process level. The process is said to be unsatisfactory if the distribution of quality characteristic is asymmetric about the specified process level. Hence, it is necessary (though not sufficient) to develop charting statistics to detect the lack of symmetry in the distribution of the process characteristic of interest. To account for such a type of changes or more precisely to account a change being from symmetry to asymmetry of the distribution of X, we need to consider classes of distributions which include both asymmetric and symmetric distributions. To be specific, a process is considered to be in statistical control if the distribution of X is symmetric and lack of this symmetry being the indicative of the process being out of control. In such situations, to identify the out of control state of a process it is necessary to develop charting statistics to detect the lack of symmetry in the distribution of the quality characteristic of interest. Further, one may use a parametric or nonparametric procedure, depending upon the form of the distribution of the characteristic of interest. One may use suitable existing univariate tests for testing symmetry about known point of symmetry, as described above. In case of multivariate process characteristic, one may use some existing parametric or nonparametric charting statistics such as Chi-square statistic, Hotelling's $T^2$-statistic, multivariate rank statistic, multivariate Wilcoxon signed-rank statistic to monitor the process.

Chapter-6 is devoted to control charts for generalized $\nu$-spherical models. Here, we assume that the X, the process characteristic(s) of interest, has generalized $\nu$-spherical density, which include a class of symmetric distributions symmetric about a specified process level. We develop some charting statistics for detecting a shift from the 'symmetry of the distribution about 0'. In Section 6.2, we develop some parametric and nonparametric charting statistics for univariate process characteristic. In Section 6.3, we develop some parametric and nonparametric charting statistics for multivariate process characteristics which follows generalized $\nu$-spherical densi-
ties. We study performance of these charting statistics together with some existing charting statistics in Section 6.4. Concluding remarks are given in Section 6.5.

In this thesis, the sections are numbered by chapter number followed by a section/subsection number. For example, Section 3.2.1 means 3rd chapter and 1st subsection of 2nd section. Further, equations, lemmas, figures and tables are numbered by starting with chapter number and a serial number. For example, Lemma 2.1 means 2nd chapter and 1st lemma. The equations are referred by such numbering within parenthesis.

Thesis ends with appendix which include bibliography contains the list of references used in this thesis. MATLAB programs are developed for computational work involved in this thesis and these are included in CD along with a thesis.