ABSTRACT

PROBLEMS RELATED TO EFFICACY MEASUREMENT AND ANALYSES

by
Sibabrata Banerjee

In clinical research it is very common to compare two treatments on the basis of an efficacy variable. More specifically, if X and Y denote the responses of patients on the two treatments A and B, respectively, the quantity P(Y>X) (which can be called the probabilistic index for the Effect Size), is of interest in clinical statistics. The objective of this study is to derive an efficacy measure that would compare two treatments more informatively and objectively compared to the earlier approaches. Kernel density estimation is a useful non-parametric method that has not been well utilized as an applied statistical tool, mainly due to its computational complexity. The current study shows that this method is robust even under correlation structures that arise during the computation of all possible differences. The kernel methods can be applied to the estimation of the ROC (Receiver Operating Characteristic) curve as well as to the implementation of non-parametric regression of ROC. The area under the ROC curve (AUC), which is exactly equal to the quantity P(Y>X), is also explored in this dissertation. The methodology used for this study is easy to generalize to other areas of application.

PROBLEMS RELATED TO EFFICACY MEASUREMENT AND ANALYSES

by
Sibabrata Banerjee

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences

Department of Mathematical Sciences

May 2007

APPROVAL PAGE

PROBLEMS RELATED TO EFFICACY MEASUREMENT AND ANALYSES

Sibabrata Baneriee

Dr. Sunil K. Dhar, Dissertation Advisor — Date
Associate Professor, Mathematical Sciences, NJIT

Dr. Manish Bhattacharjee, Committee Member — Date
Professor Mathematical Sciences, NJIT

Dr. Farid Kianifard, Committee Member — Date
Senior Associate Director, Biometrics, Novartis Pharma, NJ

Dr. Sandip Sinharay, Committee Member — Date
Research Scientist R&D, Educational Testing Service, NJ

Dr. Thomas Spencer III, Committee Member — Date
Professor, School of Management, Walden University, NJ

Dr. Kaushik Ghosh, Committee Member — Date
Assistant Professor, Mathematical Sciences, NJIT

BIOGRAPHICAL SKETCH

Author: Sibabrata Banerjee

Degree: Doctor of Philosophy

Date: May 2007

Undergraduate and Graduate Education:

- Doctor of Philosophy in Mathematical Sciences,
  New Jersey Institute of Technology, Newark, NJ, 2007
- Master of Science in Statistics,
  University of Maryland, Baltimore, MD, 2003
- Master of Statistics,
  Indian Statistical Institute, Kolkata, India, 1997
- Bachelor of Statistics,
  Indian Statistical Institute, Kolkata, India, 1995

Major: Mathematical Sciences

Presentations and Publications:

Banerjee S. and Dhar, Sunil
"Problems Related to Efficacy Measurement and Analysis"
Joint Statistical Meeting (JSM) Seattle, Washington 2006

Banerjee S. and Yoo, Wonsuk
"A Comparison Study of Models for the Human Sex Ratio,"
Joint Statistical Meeting (JSM) Minneapolis, Minnesota, 2005

Banerjee S. and Roy Anindya,
"Finite Sample Efficiency of Local linear Estimation in EXPAR Models,"
Frontiers in Applied and Computational Mathematics (FACM) Conference,
Newark, NJ 2004

To my beloved parents,
to my lovely wife,
my dear sister
and my dear parents in law

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Dr. Sunil Dhar, who introduced me to this problem. He has also given me countless ideas and suggestions. Nothing would have been possible without his constant encouragement and support. It would have been impossible to have done anything without Dr. Manish Bhattacharjee's immense help. He always had his door open. I could walk into his room with any problem and come out with a solution or a strong hint that would lead to the solution. Dr. Farid Kianifard deserves special thanks. The main approach towards solving the problem hinges on one of his papers. He has also given me a number of pointers and ideas for my research, and has followed the work very closely. Whatever I have managed to learn about the flow of a research document is mostly from him. Dr. Thomas Spencer has donated several useful books to me on mathematics and statistics from his personal collection! My sincere thanks to him. I would also like to thank Dr. Sandip Sinharay for his technical guidance from time to time. He was my senior at school and I am very fortunate to have him on my committee.

Dr. Kaushik Ghosh, was also my senior in school, and is now my professor. He has helped me immensely with his time, both with technical and non-technical issues. My wife Bhramori, being a statistician herself, helped me with her inputs from time to time. Almost everything that I have managed to learn about technical presentations is due to her.

I am also thankful to the Department of Mathematical Sciences and the University for supporting my work and for the computing and other facilities available to me.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Objective

The objective of this dissertation is to study measures of *efficacy*. Starting with the specific problem of comparing two anti-hypertensive drugs in a double-blind clinical trial, a *bootstrap kernel density estimate* of the difference of these drugs is proposed. The *bootstrap confidence intervals* of the proposed density are also computed. The study demonstrates that this method is fairly robust even under *dependent* structures that arise during the computation of all possible differences, on which some of the inference is based. A survey of the existing methods of efficacy measurement has been included highlighting the salient points and limitation of each method.

These methods are also utilized to study survival analysis, more specifically the estimates of the survival function and the hazard function. A density based algorithm to estimate the derivative of the receiver operating characteristic curve is introduced in this study. An application of this method in discriminant analysis is explored here.

This dissertation also utilizes algorithms such as *kernel density* estimation, *nearest neighbor* estimation and *ROC regression* and shows that they can be effectively adopted as applied statistical tools in the area of pharmaceutical statistics.

In the context of *kernel density* estimation, *bandwidth* selection methods play a crucial role. Literature reviews of such methods are included in reasonable detail. All the bandwidth selection methods fall under a general class of problems that can be subsumed under stochastic optimization. Stochastic optimization results and problems are the basis of classical statistical theory.

Examples of these optimization include *maximum likelihood* estimation, *likelihood ratio* test, *Neyman-Pearson Lemma* and optimization of the *bias* and the *variance* of an estimator.

1.2 Background Information

Comparing two treatments with respect to a primary efficacy variable is a problem which is commonly encountered in the clinical studies. Several parametric and non-parametric methods are used to find solutions to this problem. Parametric approaches are often based on normality assumptions. Non-parametric approaches are primarily rank based tests like the *Wilcoxon-Mann-Whitney(WMW)* test.

Effect size (ES) presents the magnitude of the difference between two treatments under consideration. In many of the recent approaches, ES is presented as a mathematical expression, not easily understood by clinicians. An ideal ES measure is somewhat easy to compute and appeals to both clinicians and statisticians. An example of such a measure is given by the probabilistic index $P(Y > X)$, where Y and X denote the performance measure of two competing drugs (one of which may be a placebo). Here it is assumed that the larger the performance value, the more efficacious is the drug.

One of the first studies that surveyed the estimation of the quantity $P(Y > X)$ is Wolfe and Hogg (1971). They have considered the estimation under the normality assumption as well as without this assumption. Confidence bounds for $P(Y > X)$ and some related quantities is also explored with several examples.

Simonoff et al. (1986) have explored the quantities $P(Y > X)$ and a reparametrized version $P(Y > X) - P(X > Y)$ in categorical data coming from two populations. In this paper they have introduced a hybrid estimator with parametric as well as a non-parametric

properties. This estimator is essentially Wilcoxon-Mann-Whitney (WMH) statistics, except the proportions are calculated using normality assumption. They have also introduced in this same paper, a smooth estimator of $P(Y>X) - P(X>Y)$ based on non-parametric density estimation procedure. They have also shown both with empirical studies and theoretical justifications by variance estimation and asymptotic results, that the hybrid estimator, *pseudo-MLE* as they call it, gives accurate results under many circumstances. Note that the smoothed estimator based on non-parametric density estimation method has been used in categorical data. Their study gives evidences that this estimator is asymptotically equivalent to the WMW statistics and bias correction methods and bootstrap treatment makes it even more attractive as a (bias reduced) smooth estimator.

Methods based on bootstrap confidence intervals of this quantity have been discussed by Chen and Kianifard (2000). It has been correctly pointed out by them that the rejection of the *Wilcoxon-Mann-Whitney* test would imply that the two distributions underlying the observations are not the same. Therefore, in an efficacy measurement analysis, it can be reasonably inferred that one treatment is better than the other. However, this does not provide a quantifiable measure of efficacy. Also, it is mentioned in the paper, and illustrated with an example that it is not reasonable to assume a shift model when the responses are categorical. The *Wilcoxon-Mann-Whitney* test is often used with a shift model assumption.

Contrast comparisons between several ES measures have been discussed by Acion et al. (2006). They have considered the commonly used 'measure of separation' between two distributions (*Cohen's d*), as one of the meaningful ES. However, the limitation of the '*Cohen's d*' lies in it's assumption of homoscedasticity, which is often not the case. The *Odds ratio* and the *generalized odds ratio* are also explored in the above mentioned paper.

The quantity $P(Y > X)$ or the probabilistic index is introduced as one of the competitive ES measures in this paper by Acion et al. (2006).

Although this probabilistic index was in use before, the close relation between the *Receiver Operating Characteristics* (ROC) curve and the quantity $P(Y > X)$ has generated renewed interest in clinical research in a paper by Acion et al. (2006). The *'Kendal's* τ*'* is also discussed in the above paper, but this is only meaningful if the data is collected in matched pairs. Brumback et al. (2006), have also developed methods of connecting the probabilistic index with the ROC and introduced the *Area under the ROC Curve* (AUC). The most involved part of their work is on accommodating for covariates. *Adjusted* AUC is an alternative index for the AUC which is computed on the basis of a semi-parametric grouping of the covariate measures which works even in difficult practical situations like confounding effects. *AUC Regression* is also mentioned in the same context. This topic is developed in detail by Dodd and Pepe (2003) and Pepe (2003).

Asymptotic approaches towards the effect size measure problems are considered by Newcombe (2006). Tail area based methods of determining efficacy is the prime focus there. Several competitive methods are discussed and a final choice is recommended on the basis of the performance comparison of these methods in the above mentioned paper.

ROC as a concept is borrowed from the signal detection theory. It is in use in the context of medical diagnostic testing for evaluation of the performance of a binary classifier with continuous output.

There is a fairly extensive literature on estimating $P(Y > X)$ where *X*, *Y* follow two unknown distributions F and G, respectively. While the interest lies in estimating this quantity $P(Y > X)$, a comparatively new way to approach this problem and draw inferences is to consider the amount of the shift D $(= Y - X)$ of the random observation

vectors X and Y. Here D itself is a random variable.

For example if it is known that X and Y are independent Normal with mean and variances μ_x, μ_y, and σ_x^2, σ_y^2, then the distribution of D would be $N(\mu_y - \mu_x, \sigma_x^2 + \sigma_y^2)$. If both of the distributions F and G are unknown, then one has little idea about D.

Nonparametric density estimation is a novel method which has been overlooked for a long time by clinical statisticians. Parametric density estimation however, is commonly used by statisticians. For example, if D is normal, one would only have to estimate the mean (μ) and the variance (σ^2) from the data. Alternately, if the density of the quantity D is Gamma, one would try to estimate the scale and the shape parameters α and λ of the gamma density.

Nonparametric density estimation is a general method to deal with unknown densities. This method works with minimal assumptions. No assumption of a specific density or even a family of distributions (location, scale, exponential family et al.), is required in this method.

This approach can be well adopted in this context. Instead of dealing with the individual densities of X and Y one can deal directly with D and look at the above mentioned quantities as the following:

$$\theta_0 = P(Y > X) = P(D > 0).$$

In the current study, non-parametric density estimation methods are applied for estimating the quantity θ_0 . Along with it, estimates of $P(D > v)$ is obtained for many threshold values v. More on this would be given in the density estimation chapter.

Apart from the non-parametric density estimation, parametric approaches also have a great potential if explored in problem specific modeling. For example exponential,

double exponential, gamma, uniform or generalized uniform assumptions lead to the corresponding *Uniformly Minimum Variance Unbiased Estimate* (*UMVUE*) of $P(Y > X)$ and inferences related to it (Ali et al. , 2005).

CHAPTER 2

NON-PARAMETRIC RESULTS

In this chapter the original problem, the study of two anti-hypertensive drugs will be described in detail. The relevance of the nonparametric density estimation methods and their application to this problem will also be pointed out. Different density estimation methods will be considered with the problem in the foreground. Survival analysis methods and receiver operating characteristic curves are very closely related to nonparametric density estimation, mainly because the same algorithms that work for estimating the density potentially helps to come up with estimates of hazard functions (failure rates), and receiver operating characteristic curves.

2.1 Defining the Problem

The following data (Table 2.1), on a double-blind clinical trial to compare two antihypertensive drugs was obtained from Hogg et al. (1990) and Chen and Kianifard (2000). The primary efficacy variable is the change in blood pressure from the baseline, i.e. difference of the blood pressure reading after the treatment with the drug from the blood pressure reading before the treatment. Most of the analyses in this kind of problem either assume normality, and involve an *unpaired* t-test, or with minimal assumptions, use a non-parametric version of this t-test, namely the Wilcoxon-Mann-Whitney (WMW) test (also called the Mann-Whitney U test).

The WMW tests are rank based tests, called the Wilcoxon rank-sum test and the Mann-Whitney U test introduced independently by Wilcoxon (1945) and Mann and Whitney (1947). The purpose of these tests is finding out if the two random variables has the same distribution, or if one of the random variables is stochastically larger than the

other. Detailed discussions of these tests along with similar rank based tests are included in Lehmann (1998). Some of the studies combine the two tests as the Wilcoxon-Mann-Whitney (WMW) test. However, the WMW test is often used where two population distributions are assumed to have the same general shape, but one of them is shifted relative to the other by a constant amount Δ, under the alternative hypothesis. The computational aspects of the WMW test statistics have been reviewed by many researchers, Bernhard et al. (1988) and Ludbrook (1995) have examined the accuracy of the WMW test in microcomputer statistical packages. Bergmann et al. (2000) have reviewed the performance of the WMW test with 11 commercial statistical packages. They have used real datasets from a pharmacological experiment and found that these commercial statistical packages gave very different outcomes from the Wilcoxon-Mann-Whitney test. The popular ones in their list of 11 packages include SAS 6.12, S-plus 2000, SPSS 8.0, JMP 3.2.5 and SYSTAT 9.

In Chen and Kianifard (2000), they have considered the efficacy measure θ_0 given by

θ_0 = $P(Y > X)$ for a continuous response variable,

= $P(Y > X) + 0.5\ P(Y = X)$ for non-continuous response variable

where Y and X are the response variables under question.

Table 2.1 Distribution of Blood Pressure Change from Baseline for Two Antihypertensive Drugs
[Source: Chen and Kianifard (2000), Hogg et al. (1990)]

Change in pressure	Treatment A Blood (frequency)	Treatment B (frequency)
-40	2	0
-29	1	0
-28	1	1
-26	2	0
-24	4	2
-22	2	0
-20	1	0
-18	2	4
-16	1	3
-15	1	1
-14	3	2
-13	1	1
-12	3	3
-10	4	5
-9	1	0
-8	7	6
-6	0	4
-5	1	0
-4	5	7
-3	0	1
-2	2	6
0	8	4
2	1	3
4	4	1
8	0	2
9	1	1
12	1	0
Total	59	59

The dataset given in Table 2.1 will be repeatedly used in the present study. A summary statistics of this dataset is included in the Table 2.2 to give a rough idea of the data, so that the reader would be prepared for the different analyses performed at a later point in this study.

Table 2.2 Summary Statistics

Summary	Max	Min	Mean	Meadian	Variance
A	12.0	-40.0	-10.2	-8.0	126.9
B	46.0	-28.0	-5.9	-6.0	125.3

It has been pointed out in Chen and Kianifard (2000), that $P(Y > X)$ is equivalent to the *Wilcoxon-Mann-Whitney* statistic in continuous data. A *bootstrap confidence interval* for the above parameter θ_0 has been computed in this paper. It has also been mentioned there that the *WMW* statistic does not give us a quantifiable measure of efficacy. Generally speaking, looking at the critical values of the Mann-Whitney U-statistic or deriving a confidence interval of the above mentioned parameter θ_0 would give us the much needed insight of the efficacy. It would also make it clear to us if the medians of the two distributions in question are significantly different, under the shift model assumption. The information that it would fail to provide would also be vital. For example it would not be able to give us information on the underlying uncertainty of the two distributions or a quantifiable difference of the means of the two distributions. Studies show that the Mann-Whitney-U statistic can be unreliable if the variances of the two populations are quite different. It also has the drawback of ignoring any information obtainable from the underlying distributions; see Simonoff (1986), Owen et al. (1964) or Birnbaum and McCarty (1958). Potthoff (1963), has introduced a modified form of the Wilcoxon statistic to test broader form of null hypothesis that involves unequal variance and absence of normality, which is commonly referred to as the Behrens-Fisher problem.

'Cohen's d' is a non-parametric measure of effect size, commonly used in psychometrics and other areas of applied statistics, and particularly in problems that have similar structure as the one cited above. It is given by the following formula

$$\frac{\bar{Y}-\bar{X}}{S_p},$$

where $\bar{Y}$ and $\bar{X}$ are the respective means and S_p is the pooled standard deviation. S_1 and S_2 are the standard deviation of the two samples of sizes $n_1 \& n_2$, respectively. The advantages of working with this measure is that it is not based on distributional assumptions and is unit free, so it is easier for non-statisticians to understand the effect size just by looking at the prescribed limits, for example, 0.2 for small effect size, 0.5 for moderate and 0.8 for high effect size (Acion et al., 2006). These limits cannot be statistically meaningful.

However, it is easy to show that with unequal variances 'Cohen's d' is not a meaningful estimator of effect size. It is easy to upset the measure and have 'Cohen's d' values smaller than 0.2 and yet have quite a separated pair of populations. For example, in the following summary statistics, the 'Cohen's d' is smaller than the prescribed 0.2, but the means are far apart:

$$\bar{Y} = \mathrm{v}\,(\bar{X})$$

$$S_1 = 23\,(S_2)$$

$$n_1 = 4n_2 = 400$$

$$\bar{X} = S_2$$

Table 2.3 'Cohen's d' Values for Different Ratios of Means of Two Samples

Ratio of $\bar{Y}$ and $\bar{X}$	Cohen's d value
1.00	0
1.50	0.024
1.75	0.036
2.00	0.049
2.25	0.061
2.50	0.073
3.50	0.121
4.00	0.146
5.00	0.194
5.25	0.206
5.50	0.219

The 'Cohen's d' value is given by 0.194 (< 0.2) and yet one can readily see that the effect size should be far more than moderate. Given in Table 2.3 are the 'Cohen's d' values for different ratios of the means, keeping the remaining relation same as the present example.

To get even a small effect size, the mean of one sample has to be at least about 5 times bigger than the other. The standard deviation of one sample being 23 times bigger than the other plays a key role in explaining the level of the effect size.

Consider the primary efficacy variable of the two drugs in question. In this case it is the change in blood-pressure after administering the drug. Note that the pair of variables may come from two different distributions, defined on the same domain, but possibly with different variance, shape and structure. That is, the probability densities of both these variables would have all the positive mass in the same subset of the real line. Additionally, they are independent of one another. The test of equality of means under these general assumptions of unequal variance is quite involved when normality is not

assumed; this is referred to as the *Behrens-Fisher Problem.*

Even if it is known that the mean or the median of one treatment is bigger than the other, one would still be unsure of the answer to the question "how much bigger ?". This question is the key focus for the measure of effect size. Consider two samples from two different populations. The observations from these two populations are independent of one another, both within and between the samples. If the study is in matched pairs, one can define D to be the vector of pair wise difference. The situation now deals with a pair of independent samples and additionally, they may not be of the same size, although in this example, the sample sizes are the same. If one could look at the vector D of all possible differences, they are draws from the distribution of the true difference. Note that they are identically distributed as F_D, the theoretical distribution of the differences, but they are not independent, as the cohort of the difference values would be correlated with each other. The density computed from the differences in the vector D would not only provide the shape of the difference, it would also provide estimates of the *median* and other *quantiles*, *mean* and the *variance* and can be utilized to get the estimate of the derivative of the density. It would also give a clear idea of the amount of separation of the two distributions. Application of *bootstrap* procedure helps in the removal of bias. At this point, it should be noted that WMW statistics itself is computed in a way that introduces the same cohort of dependence and yet the result is unbiased as it can be shown that it belongs to a larger class of symmetric statistics, namely the U-statistic.

Given a pair of vectors of observed values from the two distributions, one can look at the differences from all possible combinations. Thus, the difference D is denoted by (Y - X) which is the abbreviation for $D_{mn\times 1} = (D_{11}, D_{12}, ..., D_{mn})^T$ where Y and X are two vectors of dimensions $n\times 1$ and $m\times 1$, respectively, and $D_{ij} = Y_j - X_i$.

Mann-Whitney U is an unbiased estimator for the parameter θ_0 which is defined as

$$\theta_0 = P(Y > X) + \frac{1}{2}P(Y = X).$$

For continuous distributions, it reduces to

$$\theta_0 = P(Y > X).$$

Note that the definition of θ_0 for the non-continuous distributions involves the additional term $\frac{1}{2}P(Y = X)$. This term is introduced to adjust for ties in the sample. However, the added term $\frac{1}{2}P(Y = X)$, ensures that the information of ties is used and is consistent with the expression of incremental probability given by the following expression

$$P(Y > X) - P(X > Y) = 2\theta_0 - 1.$$

The *Receiver Operating Characteristic* curve applied in the context of efficacy measurement is also another appealing area of study, more so because of the relationship of the area under this curve and the following quantity $P(Y > X)$, the probabilistic measure of effect size.

2.2 Density Estimation Methods

Assuming that the distribution has a continuous derivative, density estimation procedures can be applied with minimal assumptions. Let the density of F_D exist. The density estimation procedure helps in visualizing the separation between the two treatment distributions. The estimate captures the shape of the true density.

Densities can be computed using various non-parametric methods such as *histogram, kernel density estimation, maximum penalized likelihood estimation,* the *nearest neighbor method* et al. Kernel methods can be based on several different kernels, with contrasting properties of the candidate kernels. Adaptive and variable kernel methods are popular in different statistical applications.

Apart from the choice of kernels, there are several methods of choice of bandwidths in *kernel density estimation.* However, different combinations of choices of bandwidth and kernels lead to estimated densities with different shapes and properties, often inherited from the kernels used for estimation. In addition, depending upon the choice of penalty functions, selection of kernel and bandwidth can be affected. Fortunately, the estimation procedure is relatively insensitive to the choice of kernels but is sensitive to the choice of bandwidth.

The most popular method of estimating densities is the histogram method. Although this method is intuitive and easy to follow, and perhaps the most widely accepted one, however, it suffers from some serious drawbacks. The most important one is the shape of this density, which can depend heavily on the chosen *bin-width* and the number of bins. Also, the histogram is just a collection of boxes stacked over one another representing the frequency or the relative frequency of the bins. However, the target density may be absolutely continuous and may have derivatives of many or all orders. Thus, a histogram density estimate may not capture some fundamental aspects of the underlying structure. The average shifted histogram (ASH) does rectify most of the problems encountered by the ordinary histogram approach. See Silverman (1986) or Scott (1992) for a detailed discussion.

The kernel density estimate, on the other hand, depends on the choice of kernels and the bandwidths. The underlying shape of the density estimate does not change much

with a reasonable choice of bandwidth. However, the wider this bandwidth, the smoother will be the shape of the density. Too small a bandwidth shows an overly rough curve, and too wide bandwidth can smooth out the important details from the shape of this density.

2.3 Survival Analysis

Hazard rate or *failure rate* is commonly used in survival analysis. It is defined as:

$$h(t)=\frac{f(t)}{1-F(t)},$$

where $f(t)$ is the density function and $F(t)$ is the cumulative distribution function. Note that if the random variables Y and X have proportional *hazard rates*, and the constant of proportionality is c, that is, if the following relation hold:

$$h_X \equiv ch_Y,$$

then ES would be equal to $\frac{c}{c+1}$. In other words

$$h_X(t)=ch_Y(t)\,(\text{for } t\geq 0)\Rightarrow P(Y>X)=\frac{c}{c+1}.$$

This is a non-parametric property. However, a similar property holds under a weaker condition, which does not assume the density.

$$\bar{F}_X \equiv \bar{F}_Y^c,$$

where F_X, F_Y are the distribution functions of the two populations and $\bar{F}_X, \bar{F}_Y$ are the corresponding survival functions, i.e. $\bar{F}_X=1-F_X$, $\bar{F}_Y=1-F_Y$.

This can also be readily seen from the following derivation

$$P(Y>X)=\int_{-\infty}^{\infty}\bar{F}_Y(t)\,dF_X(t)$$

$$= -\int_{\infty}^{-\infty} \bar{F}_Y(t)\, d\bar{F}_X(t)$$

$$= \int_{-\infty}^{\infty} \bar{F}_Y(t)\, d\bar{F}_X(t)$$

$$= \int_{-\infty}^{\infty} [\bar{F}_X(t)]^{1/c}\, d\bar{F}_X(t)$$

$$= \int_{0}^{1} u^{1/c}\, du$$

$$= \frac{u^{1/c+1}}{1+1/c}\Big|_0^1$$

$$= \frac{c}{c+1}\ .$$

Hence an inference based on the proportionality constant c is in fact directly interpretable as inferences on $P(Y > X)$. Existing reliability tests like log-rank test have been developed to draw inferences on this constant of proportionality.

2.4 Receiver Operating Characteristic Curve

Receiver Operating Characteristic (*ROC*) curve borrowed from the signal detection theory is very closely related to the *ES*. A point on the *ROC* curve can be expressed as $\left(P(Y > c \mid \bar{\mathcal{D}}), P(Y > c \mid \mathcal{D})\right)$. Where $Y > c$ indicates classification into the class of $\mathcal{D}$'s. Here $\mathcal{D}$ and $\bar{\mathcal{D}}$ denotes diseased and healthy populations, respectively. See Dodd and Pepe (2003) for a comprehensive discussion on this. The curve is constructed by joining different points based on decreasing threshold *c*, although the exact value of *c* is not represented on the curve. By definition, the range of the curve is $[0, 1]\times[0, 1]$. From the

point of view of ROC curve in the context of *ES* we would denote $P(Y > c \mid \bar{\mathcal{D}})$ as $P(X > c)$. That is, observations originally from the $\bar{\mathcal{D}}$ class are denoted as X.

The empirical *ROC* curve can be obtained from a pair of data vectors simply by plotting the empirical survival functions of the two random variables. Hence the empirical *ROC* curve may take different shapes based on different sample observations. Examples of *ROC* curves with illustrations are included in the simulation section. Given in Figure 2.1 is a plot of a ROC curve of two populations, diseased and normal, where the horizontal axis represents the normal population.

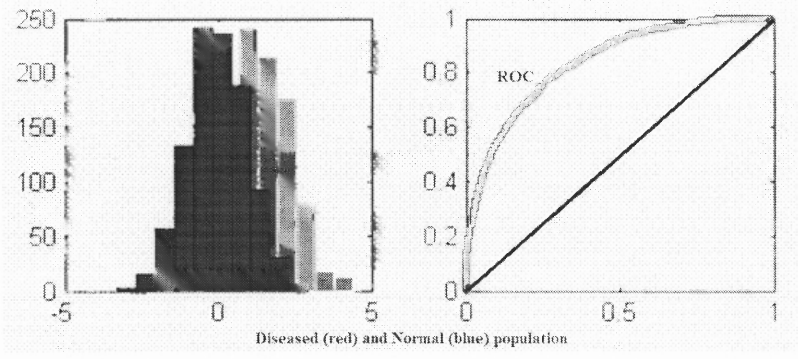

Figure 2.1 ROC curve.

Methods similar to the ones used in non-parametric density estimation can be applied to estimate the theoretical shape of the *ROC* curve. Although there is an increasing popularity of the usage of the *ROC* curve in fields like radiology and clinical statistics, not many researchers have explored the density estimation approaches in this context.

Area under the ROC curve (AUROC), as the name suggests, is the quantity obtained by measuring the area under the *ROC* curve. It is easily checked that *AUROC* is theoretically equal to the *ES*.

$$AUROC = \int \bar{F}_Y(t)\, d\bar{F}_X(t) = P(Y > X)$$

Hence, any inference based on *AUROC* is exactly equivalent to that based on *ES*. Estimates for the *ES* may also be obtained by the direct approach. This is going to be one of the primary focuses of the approaches of this study.

CHAPTER 3

PARAMETRIC RESULTS

Nonparametric ideas help us when no parametric assumptions seem to be satisfactory or even valid. But in the presence of known distributions with certain unknown parameters it is possible to find *UMVUE* of the *ES*. This problem is solved exactly the same way as one would solve a *UMVUE* problem in the parametric set up. Ali et al. (2005) have attempted to solve this problem under the assumption of Generalized Uniform distribution, and successfully found *UMVU* estimates of $\theta = P(Y > X)$ and also of $\theta^k = \{P(Y > X)\}^k$, where θ is called the reliability measure. They have discussed examples of statistical tolerance in the context of diameters of bearings and shafts.

Generalized Uniform would be a good choice in this context, but in clinical efficacy and related applications, the exponential family of distributions, particularly the exponential distribution seems to be more appropriate. Assuming a general class of distributions such as the exponential family would lead us to semi-parametric ideas, while special cases like normal distributions with known variances and exponential distribution, or the gamma distribution with the known scale parameter looks promising.

3.1 Exponential *UMVUE* for ES

Estimation of the quantity $P(Y > X)$ in the exponential distribution has been covered by many researchers. Estimation of $P(Y > X)$ where both X and Y have double exponential distribution is covered in Pal et al. (2005). In the preceding paper, double exponential distribution with a known scale parameter and unknown location parameter has been considered. The exponential distribution is a very attractive choice because of the

memoryless property and several other interesting properties such as the characterization based on absolute difference (Puri and Rubin, 1970). Sathe and Shah (1981) have explored the lower bound of the variance of the estimator of $P(Y > X)$, where Y and X are distributed exponentially with parameters λ and μ. They have also explored the bound for the Mean Squared Error of the Maximum Likelihood Estimator when one of the two parameters λ and μ is unknown.

There are several other studies on the same subject that makes it evident that estimating $P(Y > X)$ in the exponential setup is indeed a popular problem. Examples of these studies are Ivshin (1996), Cramer and Kamps (1997), Ali et al. (2004) and McCool (1991). The following result is derived assuming the exponential distribution for two random variables. The theoretical value of *ES* is denoted by θ_0.

Result. Let $X_1, X_2, ...X_m \sim \exp(\lambda_1)$ and $Y_1, Y_2, ...Y_n \sim \exp(\lambda_2)$, where the X_i's and the Y_j's are independent of each other. Then the *UMVU* for $\theta_0 \left(\theta_0 = \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)$ is given by

$$\hat{\theta}_0(s,t) = \frac{m-1}{t^{n-1}} \sum_{k=0}^{n-1} \binom{n-1}{k} (t-s)^{n-k-1} \frac{s^k}{m+k-1},$$

where $s = \sum_{i=1}^{m} X_i$ and $t = \sum_{i=1}^{n} Y_j$.

Proof: Note that $\sum_{i=1}^{m} X_i$ and $\sum_{i=1}^{n} Y_j$ are jointly *complete sufficient* for (λ_1, λ_2). So by Lehman-Scheffe theorem, as in Ali et al. (2005), the conditional expectation of any unbiased estimator of θ_0, conditioned by the *complete sufficient* statistics would be the

UMVU estimator. We will show that the following unbiased estimator of $P(Y > X)$ would serve the purpose: $I(Y_1 > X_1)$ where

$$I(Y_1 > X_1) = \begin{cases} 1 \text{ if } Y_1 > X_1 \\ 0 \text{ otherwise .} \end{cases}$$

Hence, this estimator would be given by the following expression

$$\hat{\theta}_0(\sum X_i, \sum Y_j) = E\left[I(Y_1 > X_1) \mid \sum X_i, \sum Y_j\right].$$

Denoting $\sum X_i = s$ and $\sum Y_j = t$, the above reduces to

$$\hat{\theta}_0(s,t) = P(Y_1 > X_1 \mid s,t).$$

Note that Y_1 and X_1 are conditionally independent of each other. In other words, $Y_1 \mid s,t$ is independent of $X_1 \mid s,t$, which would be clear from the following derivation

$$f(Y_1, X_1 \mid s,t) = \frac{f(Y_1, X_1, s, t)}{f_{S,T}(s,t)}$$

$$= \frac{f_{Y,T}(Y_1,t) f_{X,S}(X_1,s)}{f_S(s) f_T(t)}$$

$$= \left(\frac{f_{Y,T}(Y_1,t)}{f_T(t)}\right)\left(\frac{f_{X,S}(X_1,s)}{f_T(s)}\right). \qquad (3.1)$$

Therefore, $P(Y_1 > X_1 \mid s,t)$ can be written as

$$P(Y_1 > X_1 \mid s,t) = E\,[P(Y_1 > X_1 \mid X_1, s, t)]$$

$$= \int \bar{F}_{Y\mid t}(x) f_{X\mid s}(x) dx \,. \qquad (3.2)$$

The following formula is used to compute (3.2)

$$f(Y_1 \mid t) = \frac{f_{Y,T}(Y_1,t)}{f_T(t)} = \frac{f(t \mid Y_1) f(Y_1)}{f_T(t)} = \frac{(n-1)(t-y)^{n-2}}{t^{n-1}}, \tag{3.3}$$

using (3.3), the expression in (3.2) becomes

$$\hat{\theta}_0(s,t) = P(Y_1 > X_1 \mid s,t) = \int_0^{\psi(s,t)} \int_u^t \left(\frac{(n-1)(t-v)^{n-2}}{t^{n-1}} \, dv \right) \frac{(m-1)(s-u)^{m-2}}{s^{m-1}} \, du,$$

where $\psi(s,t) = \min(s,t)$.

$$\hat{\theta}_0(s,t) = \int_0^{\psi(s,t)} \left(\frac{t-u}{t} \right)^{n-1} \frac{(m-1)(s-u)^{m-2}}{s^{m-1}} \, du. \tag{3.4}$$

For $\psi(s,t) = s$ the above expression reduces to

$$\frac{m-1}{t^{n-1}s^{m-1}} \sum_{k=0}^{n-1} \binom{n-1}{k} (t-s)^{n-k-1} \int_0^{\psi(s,t)} (s-u)^{m+k-2} \, du.$$

By simplifying and re-arranging the terms this becomes

$$\frac{m-1}{t^{n-1}} \sum_{k=0}^{n-1} \binom{n-1}{k} (t-s)^{n-k-1} \frac{s^k}{m+k-1},$$

the UMVU estimator. Further simplification gives the following expression:

$$\hat{\theta}_0(s,t) = (m-1) \sum_{k=0}^{n-1} \binom{n-1}{k} \left(1 - \frac{s}{t}\right)^{n-k-1} \frac{(s/t)^k}{m+k-1}$$

Let $B_1 \sim Binomial\left(n-1, \frac{s}{t}\right)$. Then the above quantity is simply

$$E\left(\frac{m-1}{m+B_1-1} \right), \quad s < t.$$

Thus, it can also be viewed as the mean of a fraction, involving a binomial random variable.

However, for $\psi(s,t) = t$, the expression in (3.4) becomes

$$= \frac{m-1}{t^{n-1}s^{m-1}} \sum_k \binom{m-2}{k} (s-t)^{m-2-k} \int_0^t (t-u)^{n+k-1} du$$

$$= \frac{t}{s} \sum_k \binom{m-2}{k} \left(1-\frac{t}{s}\right)^{m-2-k} \left(\frac{t}{s}\right)^k \frac{m-1}{n+k}.$$

Thus it can also be viewed as a fraction times a ratio, involving a binomial random variable.

Let $B_2 \sim Binomial\left(m-2, \frac{t}{s}\right)$. Then the above quantity is simply

$$\frac{t}{s} E\left(\frac{m-1}{n+B_2}\right), \quad t < s.$$

So the estimator can be expressed as

$$\hat{\theta}_{UMVU}(s,t) = \begin{cases} E\left(\frac{m-1}{m+B_1-1}\right) & , s < t, \\ \frac{t}{s} E\left(\frac{m-1}{n+B_2}\right) & , t < s, \end{cases}$$

where, as mentioned before, $B_1 \sim Binomial\left(n-1, \frac{s}{t}\right)$ and $B_2 \sim Binomial\left(m-2, \frac{t}{s}\right)$.

Note that both forms of the *UMVU* estimator are valid for $m \geq 2$.

The *Maximum Likelihood Estimator (MLE)* of the above quantity $\theta_0 \left(= \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)$, using the invariance property is given by $\left(\frac{\bar{Y}}{\bar{X} + \bar{Y}} \right)$, where $\bar{X}$ and $\bar{Y}$ are sample means.

3.2 Monte Carlo Simulations of the Estimator

Table 3.1 Results from Monte Carlo Simulations

	λ_1	λ_2	$P(Y > X)$ $[\theta_0]$	$\bar{\hat{\theta}}_{MLE}$ (MSE)	$\bar{\hat{\theta}}_{UMVU}$ (MSE)
(1)	1	1	1/2	0.50078(0.00049)	0.498613(0.00049)
(2)	1	2	2/3	0.666375(0.000387)	0.665381(0.000391)
(3)	2	1	1/3	0.333192(0.000417)	0.33285(0.000481)
(4)	1	3	3/4	0.74993(0.000298)	0.74922(0.000301)
(5)	3	1	1/4	0.25087(0.0002944)	0.25032(0.0002938)
(6)	1	9	9/10	0.8995(0.000069)	0.8993(0.0000697)

To study the finite sample properties of the UMVU estimator along with the MLE estimator, Monte Carlo simulations are performed with a range of values of λ_1 and λ_2. The sizes of the samples drawn from the two exponential populations are taken to be 300 and 400. Five thousand Monte Carlo simulations are done in each case. The estimates are given along with the mean squared errors (MSE) in Table 3.1. The MSE values for the

MLE estimates are comparable to the MSE of the UMVUE. Since the MLE is easier to compute, and the UMVU estimator does not give an added advantage over the MLE as far as the MSE is concerned, it is preferable to use the simpler estimator.

CHAPTER 4

KERNEL DENSITY ESTIMATION

4.1 Approaches

Kernel density estimation can be realized as the limiting process of the *average shifted histogram* (ASH) approach. Virtually all nonparametric density estimation algorithms can be shown to be asymptotically a kernel method (Scott, 1992). The *General Kernel Theorem* due to Terrell and Scott (1992) establishes this.

Given a distribution function F, the density can be written as

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$

$$= \lim_{h \to 0} \frac{P(X \le x+h) - P(X \le x-h)}{2h} .$$

Thus for small h, the estimator of $f(x)$ is

$$\hat{f}(x) = \frac{\left\{ \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x+h) \right\} - \left\{ \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x-h) \right\}}{2h},$$

which reduces to

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} I\{X_i \in (x-h, x+h]\}}{2nh} .$$

Define $K(y) = \frac{1}{2}$ $|y| < 1$, and $K(y) = 0$, *otherwise*. Then, instead of using the indicator function in the equation above, one can use a kernel function $K(\,.\,)$ and arrive at the Kernel density estimator.

This kernel estimate of the density function can be written as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

This can also be written in the following notation, introduced by Rosenblat (1956),

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \text{ where}$$

$$K_h(w) = K(w/h)/h$$

This is the basic idea of kernel density estimation.

The *kernel density estimator* as mentioned above is also called the Parzen-Rosenblatt estimator, with a chosen *kernel function K (.)* defined as $\hat{f}(x)$ described above.

Here n is the number of observations X_1 through X_n. Although, they are supposed to be independent observations from the same distribution, the kernel method is often used when this assumption is violated. Here h is called the *bandwidth* or the *smoothing operator.* It is essentially the half length of the window in which the smoothing takes place.

For detailed descriptions of density estimation one can look at any of the following resources, Scott (1992), Hardle et al. (2004), Silverman (1986), Azzalini (1981), Seheult and Quesenberry (1971) and Izenman (1988).

The kernel density estimator is a combination from the kernels applied to each evaluation point. Thus the estimators inherit the properties of the kernels. This phenomenon is graphically expressed in the diagram (Figure 4.1). The densities given there are computed for the data on antihypertensive drugs, the example introduced in Chapter 2. The kernel functions are enlarged by a scale of ten, to enhance the visual effects.

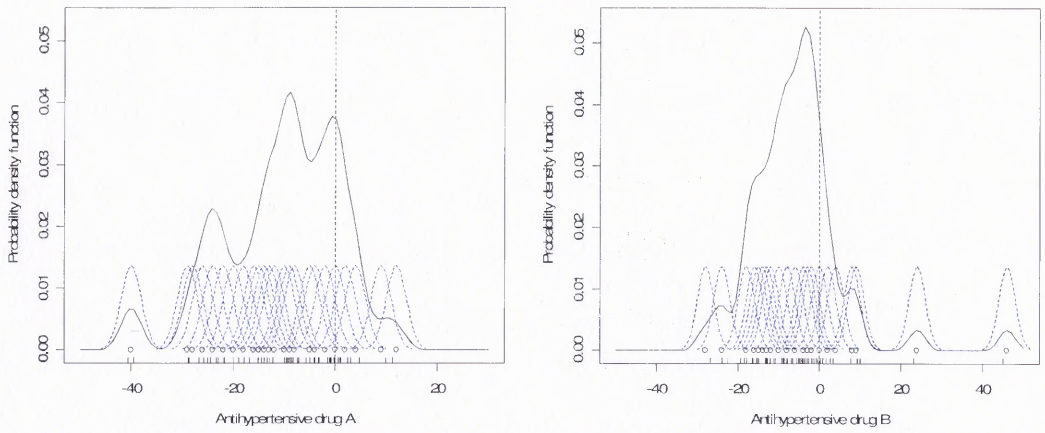

Figure 4.1 Univariate kernel density estimates of the effects of two blinded blood pressure lowering drugs along with the kernels at the evaluation points using bandwidth = 2.

In the density estimate given in Figure 4.2 a larger bandwidth is used, and that smoothed the estimate quite a bit.

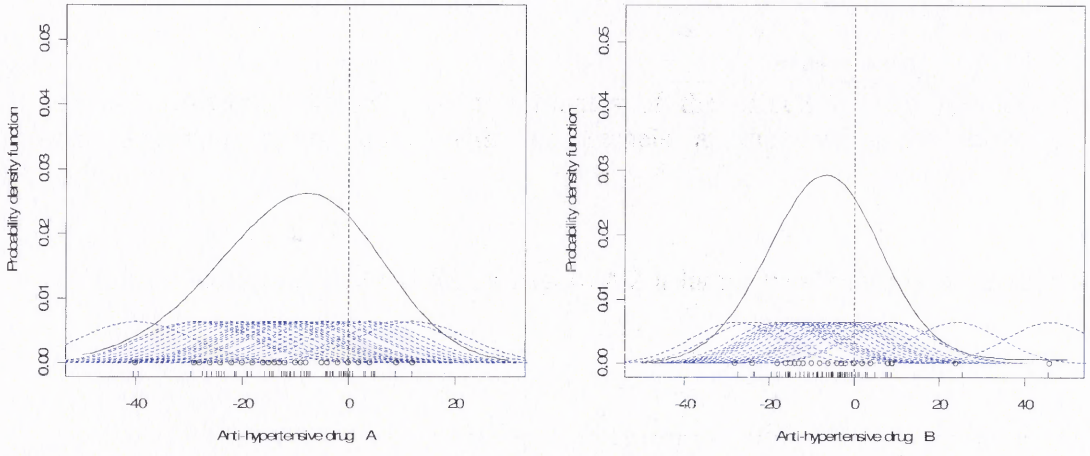

Figure 4.2 Univariate kernel density estimates of the effects of two blood pressure lowering drugs along with the kernels at the evaluation points using bandwidth = 10.

The vertical dotted line is placed on the point corresponding to zero and the curved dotted lines are the kernel functions. The density at a specific point can be thought of as a linear combination of these kernel functions evaluated at that specific point. Kernel function works as a smoother and it works like a weighted average of the data points. In histogram estimation of a density, the support is divided into several bins and

the number of observations falling into that bin is counted. Each of these observations is given equal weight and that yields the rectangular histogram. In kernel density estimation, a similar approach is taken and an interval of the length twice as much as the bandwidth is taken around a typical evaluation point. The observations falling into that interval are utilized to construct the kernel estimate and a weight is assigned to the point in such a way that the point closest to the center gets the highest weight and the point furthest from the center gets the lowest weight. This phenomenon is explained graphically in the diagram (Figure 4.2).

Some of the popular choices of kernels are given in Table 4.1. While evaluating the performance of the kernel density estimator, it is important to choose an appropriate loss function that will be able to compare the fitted density over the whole support. An expected loss function, known as the risk, can be calculated by averaging over all possible loss values. Mean Integrated Square Error (MISE) is one such choice and the approximate expression of MISE is given by $\frac{5}{4}C(K)\left(\int f^{(2)}(x)^2 dx\right)^{\frac{1}{5}} n^{-\frac{4}{5}}$ where $C(K)$ is given by $C(K) = k_2^{2/5}\left(\int K(t)^2 dt\right)^{4/5}$ and K is the chosen kernel function (Silverman, 1986). Other things being equal, the smaller the value of $C(K)$, the better the MISE will be. The expression $\left(\frac{C(K_e)}{C(K)}\right)^{5/4}$ is called the efficiency of a general kernel function compared to any arbitrary kernel function (Silverman, 1986). The symbol K_e represents the Epanechnikov kernel. It has been well established that under the assumption of independence, the Epanechnikov kernel outperforms most of the symmetric kernels as far as efficiency is considered. See Silverman (1986) for more information.

Table 4.1 Choices of Kernels

Kernels:	*Functional form*
Epanechnikov	$\frac{3}{4}(1-t^2)$
Gaussian	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$
Uniform	$I(-1 \le t \le 1)$
Triangular	$1-\lvert t\rvert \quad t>0$
Cosine arch	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}\right)t$

Although there are relatively few studies on the efficiency of kernels for dependent data, the independence assumption is often violated in real life problems and yet most of the properties of the density estimators still hold. Hall et al. (1995) have shown that even in a strongly dependent data sequence, asymptotically optimal bandwidth for independent data is a good choice, as long as some regularity conditions (such as existence of the sixth derivative of the true density) are satisfied. In this same paper they have also shown that the MISE expression for the dependent sample has a similar form as the MISE expression for the density estimate from an independent sample. A typical chart of the efficiency of other kernels compared to the Epanechnikov kernel is given in Table 4.2.

Table 4.2 Efficiencies of Kernels Compared to Epanechnikov Kernel
(Source: Silverman, 1986)

Kernels:	Efficiency
Epanechnikov	1
Gaussian	≈ 0.9512
Triangular	≈ 0.9859
Rectangular	≈ 0.9295

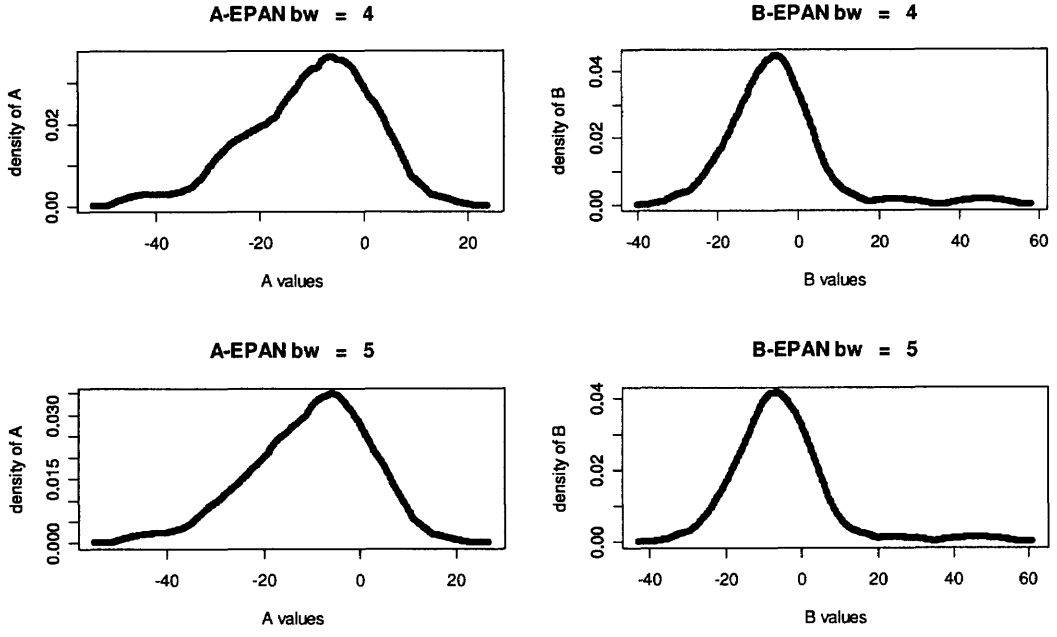

Figure 4.3 Univariate kernel density estimates of the effects of two blood pressure lowering drugs along with the Epanechnikov kernels at the evaluation points using bandwidth = 4 and 5.

Given in Figures 4.3 , 4.4 and 4.5 are the plots of univariate density estimates of the blood pressure values using different bandwidths and the Epanechnikov kernel. The bandwidth is the smoothness parameter. The larger the bandwidth is, the smoother the curve. If the chosen bandwidth is too small, it might generate spurious roughness into the density. On the other hand, if chosen bandwidth is too large, it would smooth out the curve beyond an optimal amount. Thereby it would introduce bias into the density

estimation procedure. The smoothing criterion can be illustrated using larger bandwidth with the same set of data points.

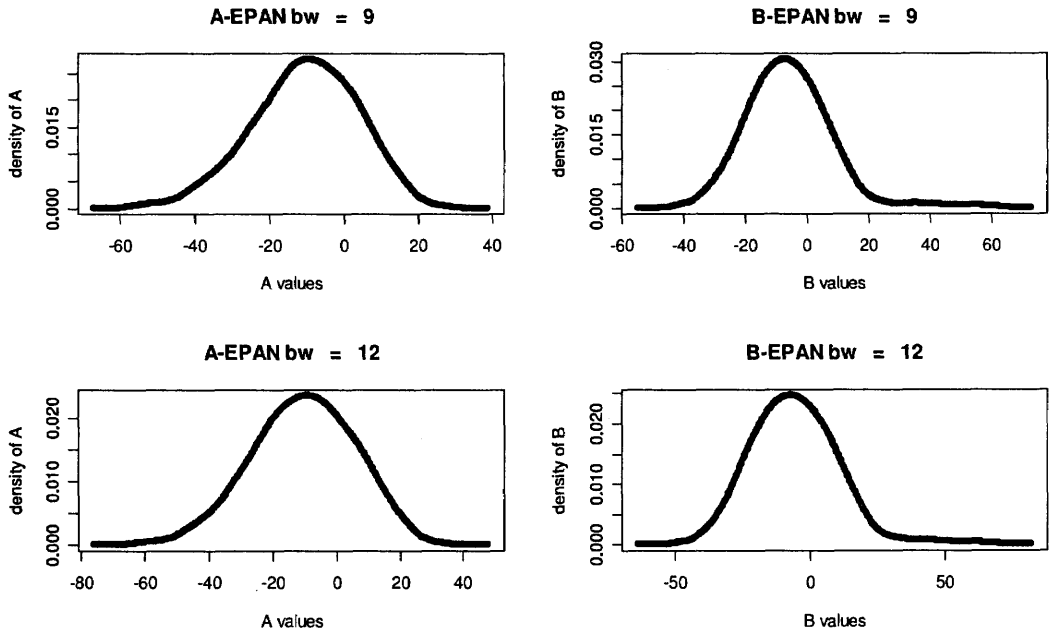

Figure 4.4 Univariate kernel density estimates of the effects of blood pressure lowering drugs along with the Epanechnikov kernels at the evaluation points using bandwidth = 9 and 12.

Consider the example in Chapter 2. The set of all possible difference from the data on blood-pressure drugs is 3481 (which is 59 squared). Even with this huge volume of data, it can be seen that with a small bandwidth, a rough spot would surface. For the same example, if densities are computed separately, the same phenomenon will surface. The illustration of this is given in the Figures 4.3, 4.4 and 4.5. The chosen bandwidths in these Figures along the first and second rows are (4, 5), (9, 12), (15, 20) respectively. That is, the first row in the Figure 4.3 corresponds to a bandwidth of 4 and the second row corresponds to bandwidth of 5. The general notion is clear from these diagrams, the density estimates become rough, smoothed and over-smoothed with the use of larger and larger bandwidth.

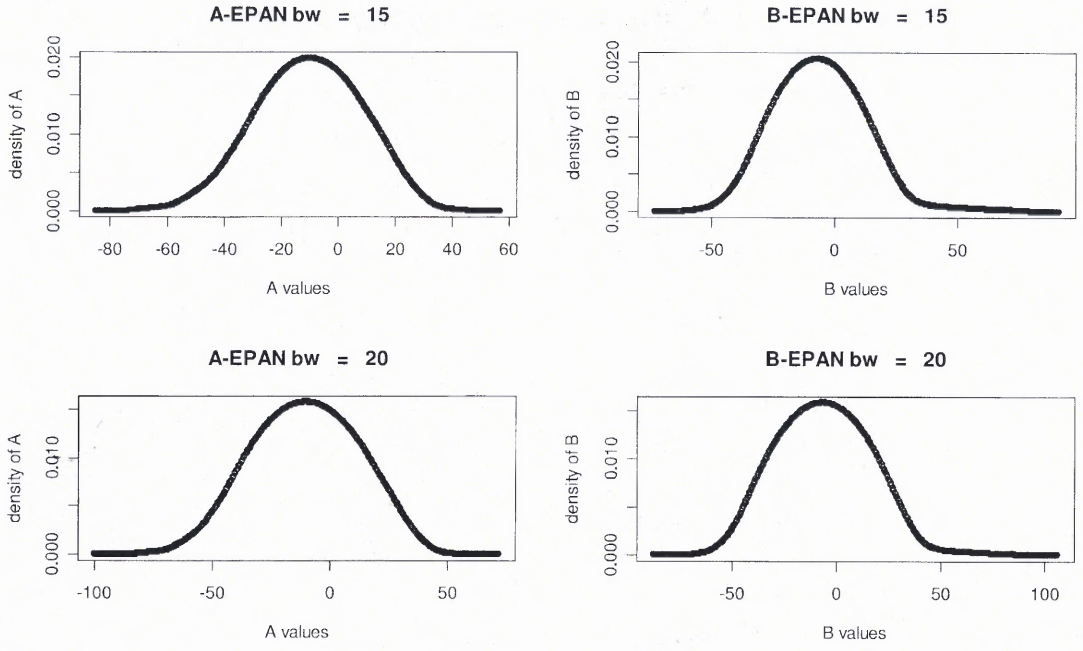

Figure 4.5 Univariate kernel density estimates of the effects of two blinded blood pressure lowering drugs along with the Epanechnikov kernels at the evaluation points using bandwidth = 15 and 20.

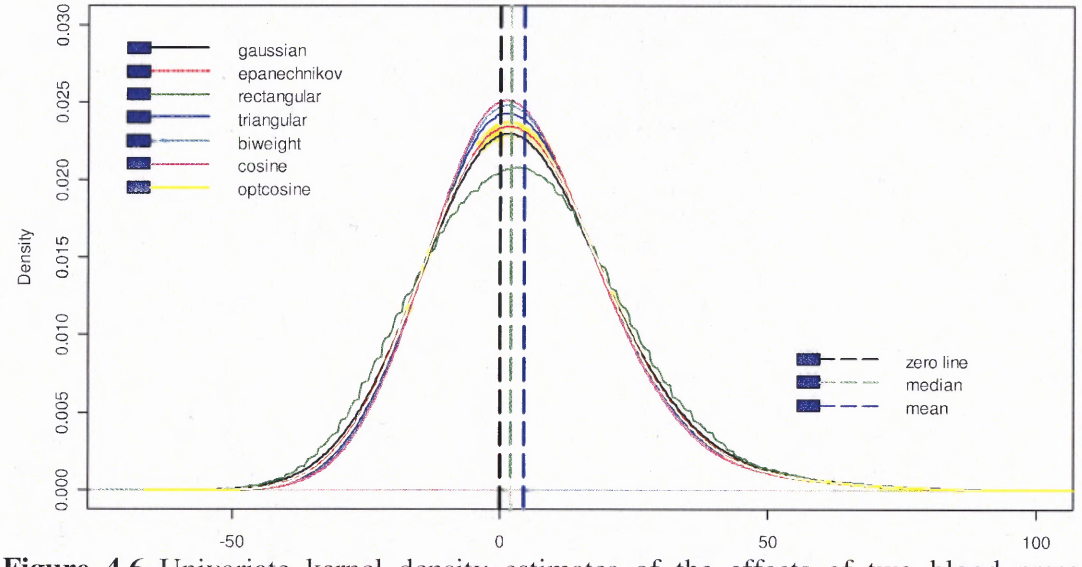

Figure 4.6 Univariate kernel density estimates of the effects of two blood pressure lowering drugs using several kernels and the same bandwidth.

In the Figure 4.6, the density plots corresponding to the kernels are color-coded for comparison. When different kernels are chosen keeping the bandwidth fixed, the estimates do not vary too much. This plot would somewhat illustrate this. Here seven

kernels are used for the smoothing operation and the estimated density of the difference is plotted in the same figure.

To investigate these density estimates with varying choice of kernels, an experiment is conducted with the same dataset, namely the variable D. density estimation results using same data and different kernels are given below. This further illustrates the fact that choice of any of these functions as kernel functions make little difference in the overall estimation method.

Table 4.3a Comparison of Density Estimates for Different Kernels Using the Same Bandwidth of 10.5

Density of D	Kernels	-24.16	3.95	22.82	42.05	70.16
Bandwidth =10.5	Gauss	0.0066 (1)	0.022 (1)	0.011 (1)	0.0026 (1)	0.00032 (1)
	Epan	0.0069 (1.048)	0.022 (0.983)	0.012 (1.017)	0.0026 (1)	0.00031 (0.984)
	Rectangl.	0.0074 (1.12)	0.022 (0.988)	0.012(1.018)	0.0027 (1.037)	0.00032 (0.987)
	Triangl.	0.0068 (1.033)	0.022 (0.99)	0.012 (1.014)	0.0026 (1.002)	0.00032 (1.011)
	Biweight	0.0068 (1.036)	0.022 (0.987)	0.011 (1.013)	0.0026 (1)	0.00032 (0.991)
	Cosine	0.0068 (1.032)	0.022 (0.989)	0.011 (1.011)	0.0026 (1)	0.00032 (0.991)
	Optcos	0.0068 (1.044)	0.022 (0.984)	0.011 (1.016)	0.0026 (1)	0.00031 (0.987)

Given in parenthesis is the ratio of value in the points of evaluation.

Table 4.3b Comparison of Density Estimates for Different Kernels Using the Same Bandwidth of 11

Density of D	Kernels	-24.91	3.65	22.81	42.35	70.91
Bandwidth =11	Gauss	0.0063 (1)	0.022 (1)	0.0114 (1)	0.0026 (1)	0.0003 (1)
	Epan	0.0067 (1.053)	0.022 (0.98)	0.0116 (1.018)	0.0026 (1.001)	0.0003 (0.984)
	Rect.	0.0067 (1.052)	0.0212 (0.964)	0.0118 (1.04)	0.00264 (0.999)	0.0003 (0.965)
	Triangl	0.0065 (1.036)	0.022 (0.987)	0.01156 (1.015)	0.00264 (1.003)	0.0003 (1.008)
	Biwgt	0.0066 (1.041)	0.022 (0.985)	0.01156 (1.015)	0.00264 (1.001)	0.0003 (0.99)
	Cosine	0.0065 (1.035)	0.022 (0.987)	0.01154 (1.013)	0.00264 (1.001)	0.0003 (0.99)
	Optcos	0.0066 (1.05)	0.022 (0.981)	0.0116 (1.018)	0.00265 (1.002)	0.0003 (0.989)

Table 4.3c Comparison of Density Estimates for Different Kernels Using the Same Bandwidth of 11.5

Density	Kernels	-24.89	3.73	22.81	42.27	70.89
Bandwidth =11.5	Gauss	0.0061 (1)	0.022 (1)	0.011 (1)	0.0027 (1)	0.0003(1)
	Epan	0.0064 (1.06)	0.021 (0.976)	0.012 (1.023)	0.0027 (1.002)	0.00029(0.985)
	Rectangl.	0.0068 (1.121)	0.021 (0.964)	0.012 (1.035)	0.0027 (0.965)	0.00029 (0.963)
	Triangl.	0.0063 (1.041)	0.021 (0.985)	0.012 (1.017)	0.0027 (1.003)	0.0003 (1.005)
	Biweight	0.0063 (1.045)	0.021 (0.982)	0.012 (1.017)	0.0027 (1.001)	0.0003 (0.989)
	Cosine	0.0063 (1.039)	0.021 (0.985)	0.012 (1.015)	0.0027 (1.001)	0.0003 (0.989)
	Optcos	0.0064 (1.057)	0.021 (0.978)	0.012(1.021)	0.0027 (1)	0.0003 (0.991)

The *kernel* function usually chosen satisfies the following properties:

$$\int K(t)dt = 1$$

$$K(t) = K(-t) \text{ or more generally } \int t\, K(t)\, dt = 0$$

$$0 < \int t^2\, K(t)\, dt < \infty$$

The asymptotic distribution of the kernel estimator regardless of the chosen kernel is given by : $N(f(x)+\beta_x, \sigma_x^2)$, where $f(x)$ is the true density and the other constants are given by:

$$\beta_x = \frac{1}{2} h^2\, f^{(2)}(x)\, \mu_2 + O(h^4) \text{ and } \mu_2 = \int t^2\, K(t)\, dt.$$

$$\sigma_x^2 = \frac{1}{nh} f(x) \int [K(t)]^2\, dt + o((nh)^{-1}).$$

Note that the *bias factor* β_x depends on the chosen *kernel* as well as the bandwidth (h) and the second derivative ($f^{(2)}$) of the true density evaluated at the point x. Therefore, minimization of *bias* would involve appropriate choice of the *kernel* and the *bandwidth.*

But this choice of the *kernel* and the *bandwidth* would affect the value of σ_x^2. Hence the goal would be to find optimal *kernel* and *bandwidth* to minimize both *bias factor* β_x and variance σ_x^2. The convention is to minimize the *Integrated Squared Error (ISE), mean integrated square error* (*MISE*) or the approximate formula for the *MISE (AMISE).*

$$\textit{Integrated Squared Error(ISE)} = \int_\Omega (\hat{f}(x) - f(x))^2\, dx$$

$$\textit{Mean Integrated Squared Error(MISE)} = \int_{\Omega} MSE \{ \hat{f}_h(x) \} dx$$

$$\textit{Approximate Mean Integrated Squared Error (AMISE)} = \frac{1}{nh} \|K\|_2^2 + \frac{h^4}{4} \{\mu_2^2\} \| f^{(2)} \|_2^2,$$

where $\| g \|_2^2$ is the squared L_2-norm, μ_2 is the second order moment with respect to the kernel function.

There are other forms of loss functions that would lead to different type of *optimal bandwidth.* One such example is the *Kullback-Leibler information* given by :

$$I(f, \hat{f}) = \int_{\Omega} \log[f(x) / \hat{f}(x)] f(x) dx$$

The optimal choice of bandwidth to minimize the *MISE* is given by :

$$h_{OPT} = \mu_2^{-2/5} \left(\int K^2(t) dt \right)^{\frac{1}{5}} \left(\int f^{(2)}(x) dx \right)^{-1/5} n^{-1/5},$$

where K is the kernel function and μ_2 is the second order moment with respect to the kernel function, and f is the true density. Different combinations of *loss functions*, *kernel functions* would give different optimal choice of *bandwidth* and hence different *density estimates*.

Taylor (1989) has explored the optimal choice of *smoothing parameter (bandwidth)* using *bootstrap*. Instead of computing the integrals for the discussed loss functions, he used the *bootstrap mean* and *bootstrap approximations* for integrals. He has shown that the optimal choice of bandwidth for this method coincides with the optimal choice in *kernel density estimation*.

Zhou and Harezlak (2002) have compared several bandwidth selection methods in *kernel smoothing* of *ROC curves*. Their empirical studies show that the *Altman smoothing method* performs best amongst the ones compared.

4.2 Estimation from Dependent Data

Kernel density estimation method is applied on data that are assumed to be independent. Dependent datasets appear naturally and therefore are often used to estimate the density. As a result, estimates that are inferior in terms of larger *variance* are obtained. If the nature of dependence is known, then an improvement can be expected in terms of lower variance, or at least a better bound of the possible bias and variance can be obtained with this extra knowledge of the dependent structure. Hence methods for reducing the bias or the variance can be adopted to further improve it. Thus, it is always important and useful to know if the dataset is dependent, and if it is what is the nature of such dependence.

When density estimates are computed from all possible differences of the two efficacy variables from independent samples, only the variance of the estimate deteriorates, and the bias remains unaffected. The asymptotic properties of the estimates such as *normality*, reduction of *bias (slow) with increased sample size*, remains to be as attractive as before. If the exact expressions of the covariances between the kernel-smoothed variables are known then one can come up with even stronger results. This is summarized and proved below.

New Result 4.2.1 Let $X_1, X_2, \ldots X_m$ and $Y_1, Y_2, \ldots Y_n$ be independent samples from two unknown distributions F_X and F_Y assuming continuous densities f_X and f_Y respectively. Let W_1, $W_2, \ldots W_M$ be the list of all possible differences of X and Y, $M=mn$, $N=m+n$. Also, if $\hat{f}_{K,h}(x) = \frac{1}{Mh}\sum_{i=1}^{M} K\left(\frac{x-W_i}{h}\right) = \frac{1}{mnh}\sum_{i=1}^{m}\sum_{j=1}^{n} K\left(\frac{x-Y_j+X_i}{h}\right)$, and $E\left(\hat{f}_{K,h}(x)\right) = \theta$, the expected value of the kernel density estimator of the dependent data at a point x, then

$$\sqrt{N}\left(\hat{f}_{K,h}(x)-\theta\right)\xrightarrow{law}N(0,\sigma^2)$$

where σ^2 is defined as : $\frac{\zeta_{10}}{\lambda}+\frac{\zeta_{01}}{1-\lambda}$, and $\zeta_{10}=\frac{1}{h^2}\mathrm{cov}\left[K\left(\frac{x-Y_1+X_1}{h}\right),K\left(\frac{x-Y_2+X_1}{h}\right)\right]$,

$\zeta_{01}=\frac{1}{h^2}\mathrm{cov}\left[K\left(\frac{x-Y_1+X_1}{h}\right),K\left(\frac{x-Y_1+X_2}{h}\right)\right]$. Where λ is equal to the constant $\frac{m}{N}$ and

thus $1-\lambda$ is $\frac{n}{N}$.

Proof Define $\varphi(X_j,Y_i) = \frac{1}{h}K\left(\frac{x-Y_i+X_j}{h}\right)$. Then the kernel density estimate

can be written as the following

$\frac{1}{Mh}\sum_i\sum_j K\left(\frac{x-Y_i+X_j}{h}\right) = \frac{1}{\binom{m}{1}\binom{n}{1}}\sum_i\sum_j\varphi\left(X_j,Y_i\right)$. Then the result follows from the

generalized Hoeffding's theorem, or the generalized U-statistics. For more detailed discussions on the generalized U-statistics see Lehman (1963).

Result 4.2.2 In the above set up, let f_w be the true density of the random variable

$Y-X=W$. Then $E\left(\hat{f}^D_{K,h}(x)\right)\approx f_w+\frac{1}{2}h^2f_w^{(2)}(x)\mu_2$. Where K is the kernel function, h is

the bandwidth used and μ_2 is $\int t^2K(t)\,dt$.

The expectation is exactly equal to the expectation of the density estimate obtained using independent data. The reason of equality is the well known fact that even in the case of correlated random variables the sum of expectation is the expectation of the sum.

$$E\left(\hat{f}^D_{K,h}(x)\right)=E\left[\frac{1}{Mh}\sum_i K\left(\frac{x-W_i}{h}\right)\right]$$

$$= 1/h\, E\left[K\left(\frac{x-W_1}{h}\right)\right].$$

(using the fact that the M random variables, although dependent, are certainly identically distributed). This can be written as the following expression

$$= 1/h \int K\left(\frac{x-w}{h}\right) dF_w(w).$$

Now let $\left(\frac{x-w}{h}\right) = -t$, hence $w = x + ht$ and $dw = hdt$. Then, the above expression becomes the following quantity (Note that $F_w(.)$ assumes continuous density f_w, which is being estimated).

$= \int K(-t)\ f(x+ht)\ dt$. Using Taylor series this is approximately equal to the following

$$\approx \int K(t)\left[f(x) + \frac{ht}{1!} f^{(1)}(x) + \frac{(ht)^2}{2!} f^{(2)}(x) + \frac{(ht)^3}{3!} f^{(3)}(x)\right] dt \quad [K(-t) = K(t)]$$

Then above would equal

$$= f(x) \int K(t)\, dt + f^{(1)}(x) \int ht\, K(t)\, dt + \frac{1}{2!} f^{(2)}(x) \int (ht)^2\, K(t)\, dt + \frac{1}{3!} f^{(3)}(x) \int (ht)^3\, K(t)\, dt$$

$$= f(x) + \frac{1}{2!} f^{(2)}(x)\, \mu_2 \quad \left[\int K(t)\, dt = 1, \int t\, K(t)\, dt = 0 = \int t^3 K(t)\, dt, \int t^2 K(t)\, dt = \mu_2\right].$$

Recall that the density estimate computed from the dependent data performs almost as well. It has all the nice root-n asymptotic properties, and normality. The bias function is also equal to the one obtained in the case of independent data. The general belief is that the bandwidth selection methods that work for independent data are not quite applicable in the case of dependent data. However, there is a relevant work by Hall et al. (1995) addressing the method of bandwidth selection in the context of density

estimation from dependent data. In this paper referred above, they have shown that even in strongly dependent data sequences, the asymptotically optimal bandwidth computed with the independence assumption is a good choice. As a matter of fact the secondary dependence induced by the kernel function applied on the dependent data structure has got stronger influence than the original dependence according to this paper.

The structure of dependence in the scenario described above can be formulated with the covariance matrix given in the following display. Let the covariance matrix of the $m \times n$ data points obtained from all possible pair wise differences be denoted by Σ.

$$\operatorname{cov}(D) = \Sigma_{mn \times mn} = \begin{bmatrix} V(D_{11}) & \operatorname{cov}(D_{11}, D_{12}) & \dots & \operatorname{cov}(D_{11}, D_{mn}) \\ \operatorname{cov}(D_{12}, D_{11}) & V(D_{12}) & \dots & \vdots \\ \vdots & \dots & \ddots & \dots \\ \operatorname{cov}(D_{mn}, D_{11}) & \dots & \dots & V(D_{mn}) \end{bmatrix}, \quad D_{mn \times 1} = \begin{bmatrix} D_{11} \\ D_{12} \\ \dots \\ D_{1n} \\ D_{21} \\ D_{22} \\ \dots \\ D_{2n} \\ \dots \\ D_{m1} \\ \dots \\ D_{mn} \end{bmatrix}.$$

Here D_{ij}'s are $Y_j - X_i$ (i = 1, 2, ..., m and j =1, 2, ..., n) and thus they are identically distributed, although they are not necessarily independent. Further denoting

$$\operatorname{cov}(D_{ij}, D_{ij'}) = \sigma_X^2 \quad j \neq j'$$

$$\operatorname{cov}(D_{ij}, D_{i'j}) = \sigma_Y^2 \quad i \neq i' \text{ and}$$

$$\operatorname{cov}(D_{ij}, D_{ij}) = \operatorname{var}(D_{ij}) = \sigma^2 = \sigma_X^2 + \sigma_Y^2 ,$$

the matrix $\Sigma_{mn \times mn}$ can be written as the following expression

$$\Sigma_{mn\times mn} = \begin{matrix} 1 \\ 2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ m \end{matrix} \begin{bmatrix} \begin{bmatrix} \sigma^2 & \cdots & \sigma_X^2 \\ \vdots & \ddots & \vdots \\ \sigma_X^2 & \cdots & \sigma^2 \end{bmatrix}_{n\times n} & \begin{bmatrix} \sigma_Y^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_Y^2 \end{bmatrix}_{n\times n} & \cdots & \begin{bmatrix} \sigma_Y^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_Y^2 \end{bmatrix}_{n\times n} \\ \begin{bmatrix} \sigma_Y^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_Y^2 \end{bmatrix}_{n\times n} & \begin{bmatrix} \sigma^2 & \cdots & \sigma_X^2 \\ \vdots & \ddots & \vdots \\ \sigma_X^2 & \cdots & \sigma^2 \end{bmatrix}_{n\times n} & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \ddots & \cdots \\ \cdot & \cdot & & \\ \begin{bmatrix} \sigma_Y^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_Y^2 \end{bmatrix}_{n\times n} & \cdots & \cdots & \begin{bmatrix} \sigma^2 & \cdots & \sigma_X^2 \\ \vdots & \ddots & \vdots \\ \sigma_X^2 & \cdots & \sigma^2 \end{bmatrix}_{n\times n} \end{bmatrix}.$$

By some simple manipulations, the above matrix $\Sigma_{mn\times mn}$ can be written as

$$\Sigma_{mn\times mn} = I \otimes P_1 \quad + \quad (J - I) \otimes P_2 .$$

Here $\otimes$ denotes the Kronecker product of matrices, I denotes the identity matrix of order $m\times m$, J denotes the $m\times m$ matrix with all the entries equal to unity, P_1 and P_2 denotes the matrices given below:

$$P_1 = \begin{bmatrix} \sigma^2 & \cdots & \sigma_X^2 \\ \vdots & \ddots & \vdots \\ \sigma_X^2 & \cdots & \sigma^2 \end{bmatrix}_{n\times n} \text{ and}$$

$$P_2 = \begin{bmatrix} \sigma_Y^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_Y^2 \end{bmatrix}_{n\times n} \text{ respectively.}$$

If the estimate of these quantities σ_X^2, σ_Y^2 and σ^2 are available, then one can easily estimate Σ.

A closed form expression of the inverse of the covariance matrix can be obtained following the result of Miller (1981). This expression in turn can be used to compute the

square root inverse of the Σ matrix and the new observations $\tilde{D}=\Sigma^{-1/2}D$ can be taken as an uncorrelated sample, since $\text{cov}(\tilde{D}_{mn\times 1})=\text{cov}(\Sigma^{-1/2}D)=\Sigma^{-1/2}\Sigma\Sigma^{-1/2}=I_{mn\times mn}$. However, this route comes to a dead end because of the following two reasons. Without assumptions about linear combination of the true density falling into the same family of density, this approach cannot be used effectively. In addition, one cannot expect uncorrelated sample to be independent. Only exception for both these restrictions is when the data are random samples from the normal distribution.

4.3 Simulations

An experiment with known densities, conducted in the context of density estimation from independent data is given below. The goal of this experiment is to see how the density estimates computed from a dependent and independent data compare with each other. Based on the theoretical results discussed in the preceding section one can expect comparable results from the two kinds of estimates. Thus estimate from a completely random sample is kept as a benchmark to judge the performance of the estimate from a dependent sample, where the dependence is artificially introduced keeping in mind the study of two drugs discussed in chapter 2. Wherever available, the true density is used in the experiment.

The independent data are sampled from a *N(1,2)* population, the dependent data are obtained from two independent samples of *N(0,1)* and *N(1,1)* population. In this simulation, the true density is *N(1,2)*, i.e. $\frac{1}{\sqrt{2}\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-1}{\sqrt{2}}\right)^2\right]$. The true density of the difference of the two normal populations *N(0,1)* and *N(1,1)* is also the same. To

cross compare the performance of the two density estimators, the true density is also used in this experiment.

Table 4.4 Comparison of Estimated Mean Integrated Squared Error of Densities Computed from Independent and Dependent Data

N	MISE (independent)	$\frac{N}{2}$	MISE (dependent)
200	0.1839	100	0.001354
210	0.06273	105	0.0003069
220	0.6591	110	0.0005882
230	0.0680	115	0.001161
240	0.1884	120	0.0007045
250	0.1688	125	0.0002284
260	0.1742	130	0.006804
270	0.2478	135	0.0007072
280	0.1119	140	0.0005643
290	0.6675	145	0.0006973
300	0.1471	150	0.0001246

The MISE along with the sample size used in the density estimations is also given. In this estimation, the Epanechnikov kernel is used and biased cross validation method is used for the selection of asymptotically optimal bandwidth. The MISE is calculated using the following expression:

$$MISE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(\hat{f}(x_i) - f(x_i)\right)^2}{f(x_i)}.$$

This method of MISE is described in Seaman et al. (1996). Here $\hat{f}(x)$ and $f(x)$ are the estimated density and the true underlying density respectively. The number n is the number of evaluation points of the densities, which is chosen to be either 512 or 1024. The range of the x values chosen for the density estimates are between -8 and 8, to ensure the inclusion of all the probable sample from each of the true underlying densities.

The independent density estimates are calculated using N number of data points where N varies from 200 to 300. On the other hand, the dependent density estimates are

calculated using all possible difference of two Normal samples, each of size 100 to 150. These numbers are exactly half of the sample sizes taken for the independent density estimate. Thus there are $\frac{N^2}{4}$ number of dependent data points and just like in the previous case, N varies from 200 to 300. Interestingly, it is observed that the density estimates corresponding to the dependent data are by far a lot better than their independent counterpart. The Table 4.4 illustrates this finding.

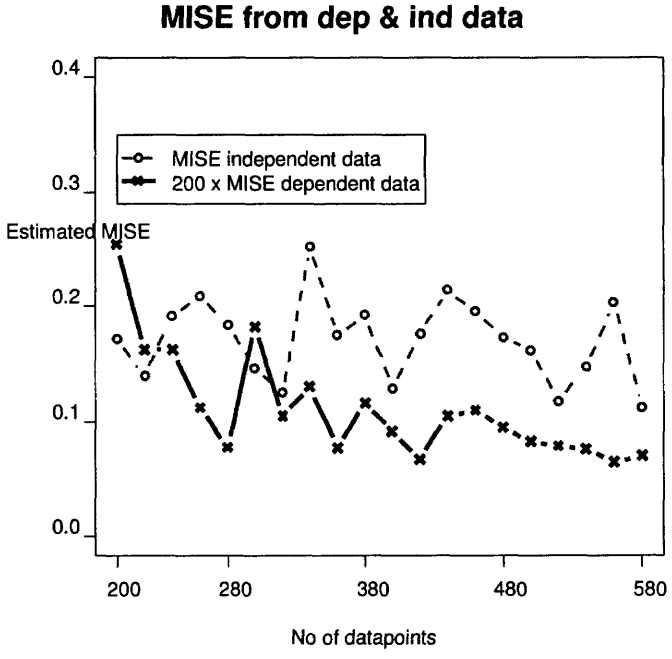

Figure 4.7 Mean integrated squared error (MISE) comparison of the density estimates computed from an independent sample and a dependent sample.

The MISE values computed from density estimates based on independent data, along with their counterpart (density computed from dependent data) are given in Figure 4.7. To keep them on the same scale, the MISE from dependent data is multiplied by a factor of 200. Yet the pair-wise MISE values for sample sizes 200 through 580 seem to be comparable.

Hence if the cost of sampling is a constant, then the simulation clearly shows that the density estimation from the dependent dataset performs better than the independent set, because in both the cases we sample the same number of variables, N and $2\times\frac{N}{2}$.

In Figure 4.8 a single instance of the density estimate for $N = 250$ is given along with the true density and the dependent density computed from the pair of normal sample each of size 125.

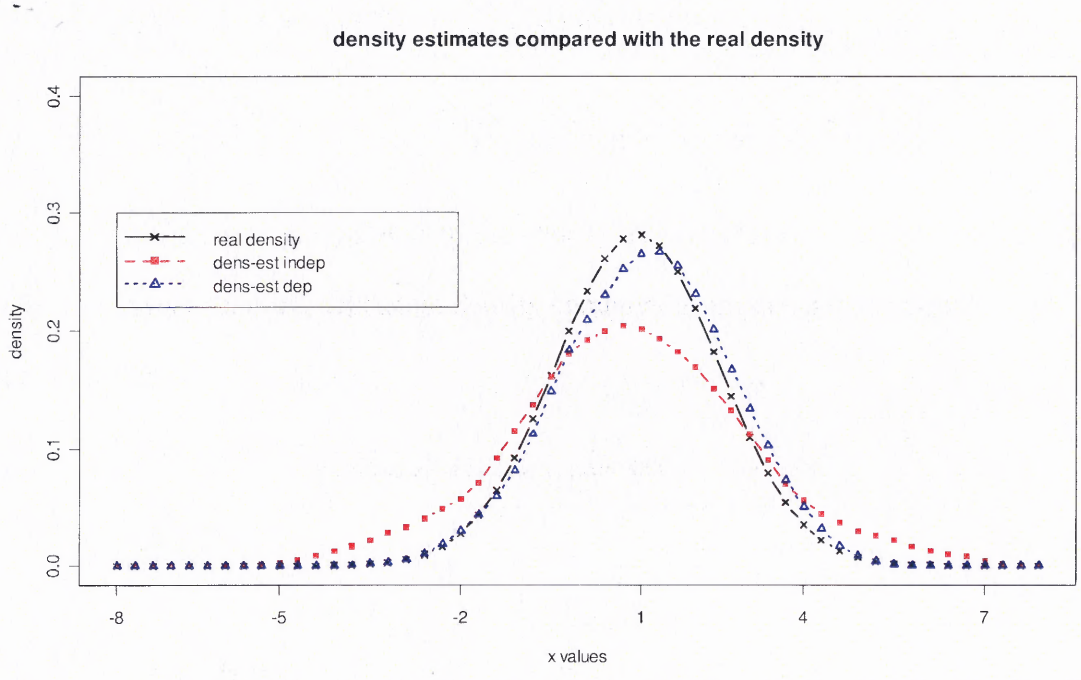

Figure 4.8 Density estimates compared with the true density.

However, one can correctly point out that the estimates computed in the case of independent dataset are based on N data points, whereas there are $\frac{N}{2}\times\frac{N}{2}=\frac{N^2}{4}$ data points (a far larger number), used in the estimation from the dependent set. Hence, overall precision in the dependent case might be lucrative but the 'per unit precision' may not be comparable. Interestingly, even the following experiment shows quite satisfactory result.

We sample $\frac{N}{2}, \frac{N}{2}$ and $\frac{N^2}{4}$ units from *N(0,1)* , *N(1,1)* and *N(1,2)* population, create the dependent sample of size $\frac{N^2}{4}$ and compare the density estimated from these with the estimate obtained from the independent sample and the actual density *N(1,2)*.

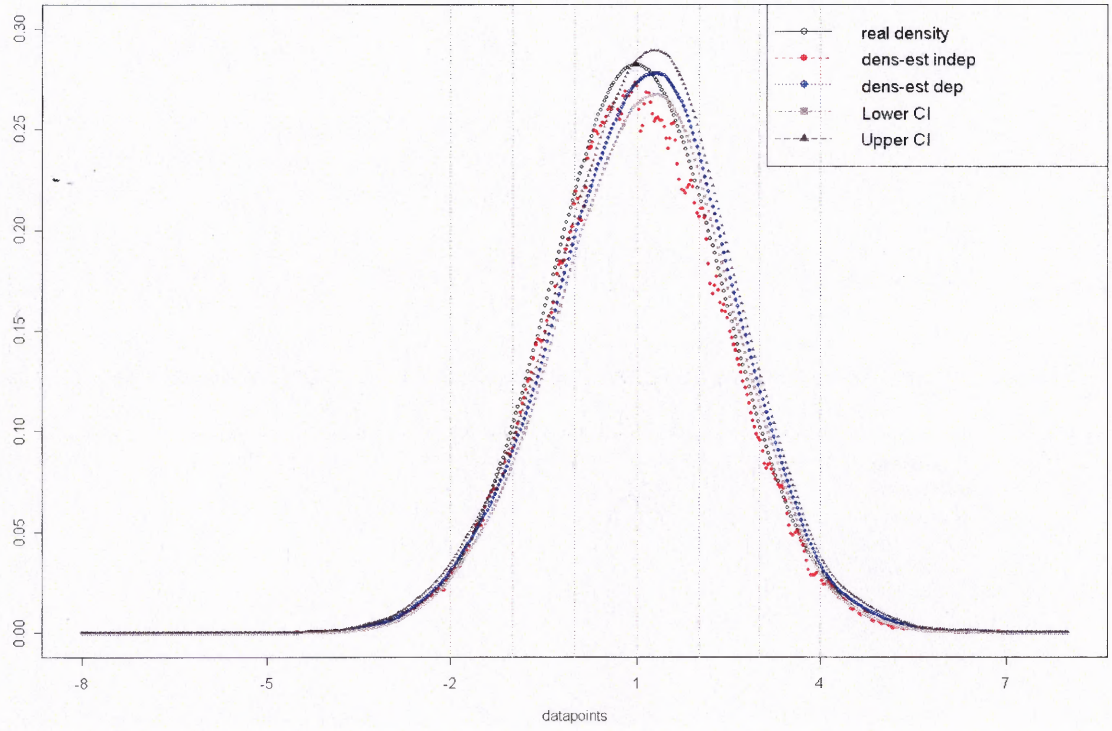

Figure 4.9 Density estimates from a very large number (40,000) of independent observations.

In Figure 4.9, the reported density estimate from the dependent data is computed from a pair of independent samples of size 200. Thus the artificially generated sample has a size 40,000. This is exactly equal to the size of the independent sample drawn from the *N(1,2)* population. Note that both the density estimates computed from dependent and independent data performs very good.

Recall that the density of D (difference between two drugs with respect to the primary efficacy variable) has been computed from the dependent data. It has also been mentioned that this density has the point-wise root-n consistency and a bias that is

unaffected by the particular structure of this dependence. However, the ultimate goal is to utilize this as an effect size measure.

Table 4.5 Estimated Probabilities Along with the Mean Squared Errors

Values(ν)	P(D>ν)	MSE
-40	0.99976	3.00E-08
-35	0.99875	2.30E-07
-30	0.99484	9.00E-07
-25	0.98341	3.33E-06
-20	0.95832	8.32E-06
-15	0.90722	1.78E-05
-10	0.82495	3.24E-05
-5	0.70766	4.74E-05
0	0.56744	5.93E-05
5	0.43804	5.89E-05
10	0.31289	5.39E-05
15	0.21579	4.23E-05
20	0.14384	2.99E-05
25	0.08993	2.09E-05
30	0.05858	1.41E-05
35	0.03922	9.43E-06
40	0.02598	6.52E-06

Given in Table 4.5 are the different possible values of the difference (D) and the estimated probabilities for the data set in Table 2.1 in Chapter 2. The mean squared errors based on bootstrap are also given along with these values.

To fix ideas, note that the estimated probability of the difference being larger than 0 is given by 0.56744 (5.931e-05), where the quantity given in parenthesis is the bootstrap based mean squared error of this probability. For a randomly chosen subject, the probability of observing a favorable response under the application of treatment B relative to treatment A is 0.56744, where the similar quantity based on Mann-Whitney statistics is 0.57541 (Chen and Kianifard, 2000).

Table 4.6 Estimated Probabilities Given Along with 95% Bootstrap Confidence Intervals and Empirical Probabilities

Values(τ)	Lower CI	P(D>τ)	Upper CI	Empirical
-40	0.999368491	0.99976	0.999983637	0.99971273
-35	0.997702871	0.99875	0.999574249	0.99885090
-30	0.992826238	0.99484	0.996574678	0.99569089
-25	0.97959232	0.98341	0.986837100	0.98592359
-20	0.952530714	0.95832	0.963829930	0.95834530
-15	0.898478238	0.90722	0.915346210	0.91209423
-10	0.813357632	0.82495	0.836072400	0.81901752
-5	0.694282778	0.70766	0.720964500	0.71760988
0	0.552225927	0.56744	0.582314300	0.54869290
5	0.422402798	0.43804	0.452674400	0.43263430
10	0.298386418	0.31289	0.327014300	0.29187015
15	0.202989419	0.21579	0.228404300	0.21344441
20	0.133045898	0.14384	0.154516900	0.13329503
25	0.081001479	0.08993	0.098868600	0.08532031
30	0.051347632	0.05858	0.065931500	0.05630566
35	0.033254946	0.03922	0.045351700	0.03964378
40	0.021122538	0.02598	0.031110700	0.02470554

Similarly one can look at several possible values of this difference and the corresponding probability along with its mean squared error. Therefore this approach gives a more elaborate understanding of the difference D, and the corresponding bootstrap based 95% confidence interval is given in Table 4.5.

Compared to Chen and Kianifard (2000), this approach of effect size measure has some fundamental difference. In the paper mentioned above, the blood-pressure variable is assumed to be discrete, but in the present study it is assumed to be continuous (otherwise the existence of the density cannot be assumed), although the observations are discrete.

The point-wise 95% bootstrap confidence interval of the probability of the difference value greater than zero is sharper than the confidence interval given in Chen and Kianifard (2000). Also, based on this study, the corresponding confidence interval of

the incremental probability of observing a more favorable response from treatment B relative to treatment A, that is $P(Y > X) - P(X > Y)$, is (0.1044519, 0.1646286), which does not include 0. Thereby it assures that the difference is significant at the 5% level.

CHAPTER 5

RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

5.1 Using ROC

Receiver Operating Characteristic (ROC) curve is a plot of 'sensitivity' (correct assertion of disease) versus '1 – specificity' (incorrect positive) of a test. In other words, the probability of a diseased subject being correctly identified is 'sensitivity' and the probability of a subject without the disease being correctly identified as healthy is 'specificity'. This curve is plotted with respect to the thresholds of discrimination of the test under discussion, where 'sensitivity' is represented along the vertical axis and '1-specificity' is represented along the horizontal axis. The concept of ROC has originated in the signal detection theory but is rapidly gaining acceptance in medical applications, several branches of engineering and psychometric applications.

References to ROC can be found even in classical regression problem and discriminant analysis, machine learning and data mining. The following section discusses some estimation problems in this area.

5.2 ROC from Density

Recall that the ROC curve is the plot of 'sensitivity' versus '1 – specificity' with respect to several thresholds of discrimination. If the observations from the diseased group is denoted as Y and the observations from the healthy group denoted by X, then the standard way of interpreting the ROC curve is $R(p) = \bar{F}_Y(F_X^{-1}(1-p))$, where the distribution of Y and X are F_Y and F_X respectively. The point p is the survival probability of X at a specified point. Also the survival function of Y, i.e. $(1-F_Y)$ is denoted by $\bar{F}_Y$ and the

inverse of the distribution function of X is denoted as F_X^{-1}. Assuming that the cumulative distribution function is strictly monotone, the value of F_X^{-1} is unique. Otherwise, if the strict monotonic property is relaxed then F_X^{-1} represents the generalized inverse of the cumulative distribution function. The functional form of the generalized inverse is given by the following expression

$$F_X^{-1}(p) = \inf \{ x \mid F_X(x) \geq p \}.$$

Obtaining the inverse of a distribution of a function is often quite complicated from a theoretical point of view, and is expensive from a computational point of view. Another way of representing the ROC curve is by using the following notation $F_X^{-1}(1-p) = c$. Hence $1 - F_X(c) \equiv \bar{F}_X(c) = p$. Thus the same ROC curve plotted as $R(p)$ against p can be seen as $R(p) = \bar{F}_Y(F_X^{-1}(1-p)) = \bar{F}_Y(c)$ versus $\bar{F}_X(c)$. In other words, the above reparametrization helps us identify the ROC curve as the plots of the survival functions of the variables X and Y, where the former represents the healthy group of subjects and the latter represents the diseased group of subjects.

An important use of ROC curve is to evaluate performances of tests that have a binary decision space and are based on continuous output. If the ROC curve based on 'test A' goes nearer to the (0, 1) point, compared to ROC curve based on another test, 'test B', then 'test A' would be a better test compared to 'test B'. Along the same lines, a gold standard test would have a ROC curve going through a point that is very near to the (0, 1) point, which would be denoted as the *ideal point* from now on. The ROC curve displays the possible trade-offs between the two kinds of errors (Pepe, 1997). So another important use of the ROC curve is to determine the optimal threshold point in order to maximize both 'sensitivity' and 'specificity'. This is same as maximizing 'sensitivity'

and minimizing '1 – specificity'.

Note that the (0, 1) point corresponds to 100% sensitivity and 100% specificity. Pepe (2003) has detailed discussion on the statistical properties of ROC curve. However, if there is a test that would randomly classify the subjects into diseased and healthy group, then the ROC curve based on that test would be the line of equality, represented by $\bar{F}_X(c) = \bar{F}_Y(c) \ \forall c$.

The choice of the optimal threshold value that should be used would depend on many factors such as the associated cost, prevalence of the disease and so on (Pepe, 1997). In the absence of the cost information, one way to choose the optimal point is to make it closest to the *ideal point* (van Belle, 2002).

An alternate way an optimal threshold value can be chosen is by choosing the point as far as possible from the line of equality, which is the ROC of the ad-hoc test, the test that randomly classifies diseased subjects from the healthy ones. In the following section, an estimation method of the optimal threshold point would be developed using the density estimation approach. The results 5.2.1 and 5.2.2 would be used for the estimation algorithm.

Result 5.2.1 $R'(p) = \dfrac{f_Y(c)}{f_X(c)}$, where $F_X^{-1}(1-p) = c$.

Proof: Note that $R'(p) = \dfrac{\partial R(p)}{\partial p}$, which can be written as the following

$= \dfrac{\partial \bar{F}_Y(F_X^{-1}(1-p))}{\partial p}$, which is essentially $\dfrac{\partial \bar{F}_Y(c)}{\partial p}$, and thus reduces to

$$= \frac{\partial \bar{F}_Y(c)}{\partial c} \times \frac{\partial c}{\partial p}$$

$$= -f_Y(c) \times \frac{\partial c}{\partial p}$$

Now $p = \bar{F}_X(c) = 1 - F_X(c)$. Thus, $\frac{\partial c}{\partial p} = \frac{1}{\frac{\partial p}{\partial c}} = \frac{1}{\left(\frac{\partial(1-F_X(c))}{\partial c}\right)} = \frac{1}{-f_X(c)}$.

Hence $-f_X(c) \times \frac{\partial c}{\partial p} = -f_Y(c) \times \frac{1}{-f_X(c)}$ (assuming that $f_X(c)$ is not equal to zero).

Thus $R'(p)$ becomes equal to

$$-f_Y(c) \times \frac{1}{-f_X(c)} = \frac{f_Y(c)}{f_X(c)}.$$

The above result was stated without proof in Lloyd (2002).

This simple but useful result can be utilized to obtain the derivative of the ROC curve at different threshold points. A direct use of the above result is to find the optimal threshold point based on the maximal vertical distance from the ad-hoc test. Figure 5.1 is an oversimplified picture of a ROC curve, and the point R on the ROC curve is the point furthest from the line joining (0, 0) and (1, 1).

Result 5.2.2 The optimal threshold point of discrimination based on the furthermost from the ad-hoc test is given by $\arg\min_c \left|R'(p(c)) - 1\right|$.

Proof: A general diagram of our problem is given in Figure 5.1. Any point on the ROC curve is given by $R(p)$. Let the perpendicular distance of an arbitrary point on the ROC curve from the line of equality be *H(p)*.

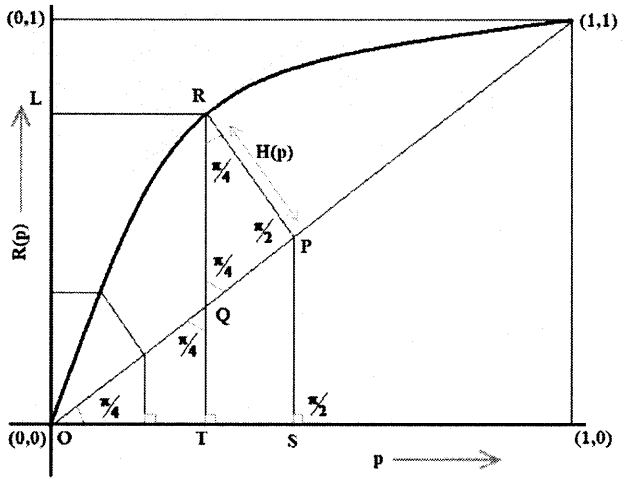

Figure 5.1 Finding the optimal point of discrimination using the ROC curve.

Then it is easy to see that $H(p) = \frac{R(p)-p}{\sqrt{2}}$. The line of equality (set of points $[p,\ p]$) intersects the perpendicular from the $R(p)$ curve into two parts, namely p and $R(p)-p$. Consider the point R on the ROC diagram. $QT = p$ and $QR = R(p)-p$. QP and $QP = RP$ are equal. So $2[H(p)]^2 = [R(p)-p]^2$ and $H(p) = \frac{R(p)-p}{\sqrt{2}}$. $H'(p) = \frac{R'(p)-1}{\sqrt{2}}$, and $H^{(2)}(p) = \frac{R^{(2)}(p)}{\sqrt{2}}$. Equating $H'(p) = 0$ and noting the fact that $R(p)$ is a concave function, its second derivative is negative; thereby it assures a point of maxima. Since there is only one extrema it is the global one. So this maxima corresponds to $R'(p) = 1$. Hence $|R'(p(c))-1|$ reaches the minimum value (0) at this point. Call this point p^* and the corresponding c value as c^*.

Finding the optimal point for discrimination with respect to the two kinds of possible errors is a common problem. In logistic regression it is common to come across a similar problem. For example let a logistic regression equation be given as the following

$$\text{logit}(\hat{\pi}) = X^T\hat{\beta}$$

Although the π_i values can be estimated by fitting the logistic regression curve, these are essentially the means of the original y_i's. So it remains to predict or estimate the possible y_i values from the obtained estimates of the π_i values. This gives rise to a discrimination problem and although choosing in the following way is the convention, this is not necessarily the best choice of π_0.

$$y_i = \begin{cases} 1 & \text{if } \pi_i > \pi_0 = 0.5 \\ 0 & \text{otherwise} \end{cases}$$

For more details, see Agresti (2002).

Recall that the optimal point of separation is c^* and the corresponding value of optimal p is given by p^*. The following diagram suggests that $\left[\frac{R(p^*) - p^* + 1}{2}\right]$ provides a lower bound of the area under the ROC curve.

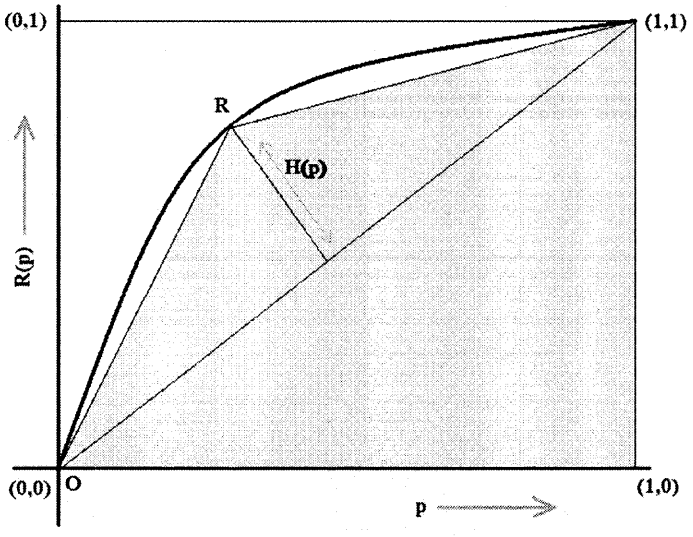

Figure 5.2 Area under the ROC curve.

The functional value of the quantity $H(p)$ is $\frac{R(p)-p}{\sqrt{2}}$. Therefore, the area of the triangle drawn by the points (0,0), R and (1,1) is given by $\frac{R(p^*)-p^*}{2}$. Hence the total area of this triangle and the triangle formed by (0, 0), (1, 1) and (1, 0) would be $\left[\frac{R(p^*)-p^*+1}{2}\right]$, which is trivially a lower bound for the area under the ROC curve. Note that the ROC curve is assumed to be concave, and in a practical situation this may not be the case. In ROC, the stochastically larger variable $\left(P(Y>c) \geq P(X>c), \forall c\right)$ is usually represented along the vertical axis. However, a concave ROC can be flipped, if the stochastically larger variable is represented along the horizontal axis. In the original form of the ROC curve, the represented test has to be at least as good as the ad-hoc test, therefore the curve stays above the line of equality.

5.3 Estimating the Derivative of ROC and the AUC Using the Density

The following algorithm would be useful to estimate the value of c^*, the optimal threshold value. Hence if there is a sample of mixed observations from f and g then their classification into one group or the other can be done if an individual observation falls on the right side or the left side of c^*.

Step 1 Choose an interval of threshold points $\mathbb{C}$, such that this set excludes the points where either of the density values may be zero. Clearly the optimal threshold point if exists would belong to this set $\mathbb{C}$.

Step 2 Choose a set of values $\{c_1, c_2, ...c_n\} \in \mathbb{C}$, for a reasonably large value of n.

Step 3 Corresponding to each c_i compute p_i.

Step 4 Look at the set of ratios $\{r_1, r_2, ...r_n\}$, where $r_i = \frac{f_Y(c_i)}{f_X(c_i)}$.

Step 5 Find the optimal value c^*, which corresponds to $r_i = 1$.

Step 6 Since the corresponsing p's are already known, it would be easy to get back to these original (X, Y) values.

To find a reasonable error bound for the optimal c^*, one can use $\varepsilon > 0, \varepsilon << 1$ such that $\overline{C^*}$ is the set containing c's for which the r value is between $[1-\varepsilon, 1+\varepsilon]$. Figure 5.3 would more clearly illustrate the optimal property of the threshold value c* and also the fact that the ratio of the density is equal to one at this threshold point.

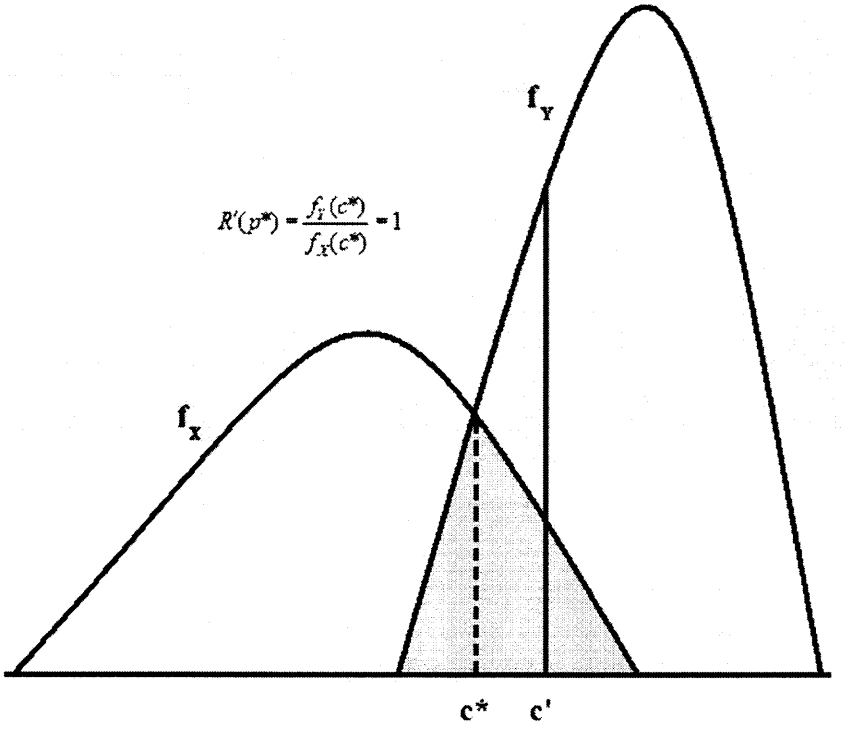

Figure 5.3 The optimal threshold value c*.

However, the algorithm described in this section guarantees the existence of the optimal point with an exception of the trivial cases of completely separated densities or

almost completely identical densities. The uniqueness can be shown for unimodal distributions with concave ROC curve. If the ROC curve has waves, then the optimal point can be chosen from several candidate points in way which is analogous to choosing a global extrema from the candidate local extremas.

5.4 ROC as a Measure of Efficacy

A reparametrized version of the ROC curve is the plot of the two survival functions of the diseased and the healthy distributions, $[P(Y > c) \mid \text{ঔ}]$ and $[P(Y > c) \mid \bar{\text{ঔ}}]$, computed for decreasing threshold values. This version of the curve can be used in the study of effect size. Following Brumback et al. (2006), one can replace the continuous outcomes of the diseased population $(Y \mid \text{ঔ})$ and the healthy population $(Y \mid \bar{\text{ঔ}})$ by Y and X respectively, where Y and X denote values of the primary efficacy variable for the drugs under study. The ROC curve in this context would appear to be closer to the *ideal point* if the outcome of the more efficacious drug is represented along the vertical axis, assuming that higher value of the outcome would correspond to better efficacy.

Since the survival function of the two underlying variables are computed at several possible thresholds c, one can compare the corresponding probabilities, at these threshold values. On the other hand one can summarize the whole ROC function into a single value, namely the area under the curve (AUC), which can be shown to be equal to $P(Y > X)$. So the ROC curve gives the comparison of the values $P(Y > c)$ and $P(X > c)$ for different values of c and the area under it is an estimate of the quantity $P(Y > X)$.

A function closely related to the reparametrized version of the ROC curve is the vertical shift function (Λ) given by

$$\Lambda(p) = F_Y \circ F_X^{-1}(p) - p,\ 0 \le p \le 1,$$

where $(Y \mid \mathcal{D})$ has the distribution F_Y and $(Y \mid \bar{\mathcal{D}})$ has the distribution F_X. Here $\mathcal{D}$ and $\bar{\mathcal{D}}$ denotes the two underlying properties of the populations being considered. For example $\mathcal{D}$ could be the placebo group and $\bar{\mathcal{D}}$, the group where the drug under study has been administered. The relation between the ROC curve and the vertical shift function defined above is given in Ghosh and Tiwari (2007)

$$R(p) = p - \Lambda(1-p),\ 0 \le p \le 1.$$

Here $R(p)$ is the ROC curve, and the function $\Lambda(.)$ denotes the vertical shift function. For some early development of the shift functions in the context of two sample problems see Docksum (1974).

Hence, the shift function can also be utilized in the effect size estimation. For example, this can help in the estimation of the area under the ROC curve under different threshold values. This in turn would be useful to identify the superior drug when there are two drugs under study.

The following two results 5.4.1 and 5.4.2 are special cases of the results by Ghosh and Tiwari (2007). Their result is based on *Generalized Rank Set Samples (GRSS)* and in the present context is simplified for the case of *Simple Random Sample (SRS)*. These two results given below can be effectively used for the estimation of the point-wise variance of the ROC curve and then utilized to create a confidence band for the ROC curve. This can also be extended in finding a confidence interval for the ES obtained as the area under the ROC curve.

Result 5.4.1 If two *Simple Random Samples* (SRS) of size m and n are collected from two populations, where the population ROC curve is denoted by R, then the empirical ROC curve given by $\hat{R}$ has the following asymptotic property

$$\sqrt{m+n}(\hat{R}-R) \to \mathbb{Z}_{R,SRS} \ ,$$

where $\mathbb{Z}_{R,SRS}$ is a zero-mean Gaussian process with covariance kernel $K_{R,SRS}(p,q)$ given by

$$\frac{f_Y \circ F_X^{-1}(p) \times f_Y \circ F_X^{-1}(q)}{f_X \circ F_X^{-1}(p) \times f_X \circ F_X^{-1}(q)} \times \frac{(p \wedge q - pq)}{\lambda} + \frac{F_Y \circ F_X^{-1}(p \wedge q) - F_Y \circ F_X^{-1}(p) \times F_Y \circ F_X^{-1}(q)}{1-\lambda}.$$

Here p,q lies in the open interval $(0,1)$, $p \wedge q$ denotes the minimum of p and q. f_X and f_Y are the densities corresponding to the distributions F_Y and F_X.

Result 5.4.2 Under the conditions of the above result, the sample variance of the function $\sqrt{m+n}(\hat{R}-R)$ at a point $(1-p)$ is given by the following function

$$\frac{\left(f_Y \circ F_X^{-1}(p)\right)^2}{\left(f_X \circ F_X^{-1}(p)\right)^2} \times \frac{(p-p^2)}{\lambda} + \frac{\left(F_Y \circ F_X^{-1}(p) - \left(F_Y \circ F_X^{-1}(p)\right)^2\right)}{1-\lambda}, \qquad (5.1)$$

where p lies in the open interval $(0, 1)$.

Therefore if we call the expression in (5.1) as *K(p)*, then a point-wise confidence band of the ROC curve can be constructed inverting the above relations. This confidence interval is given in the following equation

$$P\left(\hat{R}(p) - \frac{z_{\alpha/2}\sqrt{K(p)}}{\sqrt{m+n}} \le R(p) \le \hat{R}(p) - \frac{z_{1-\alpha/2}\sqrt{K(p)}}{\sqrt{m+n}}\right) = 1-\alpha \qquad (5.2)$$

where p is the evaluation point, $R(p)$ is the true ROC curve evaluated at p, $z_{\alpha/2}$ is the standard normal percentile with area to the right equal to $\alpha/2$.

Note that the empirical ROC curve ($\hat{R}$) mentioned above is computed from the raw data just like the empirical distribution is computed. The empirical ROC of the example of the double-blind drugs introduced in chapter 2 is given in the following

Figure 5.4. The horizontal axis represents the drug A and the vertical axis represents the drug B.

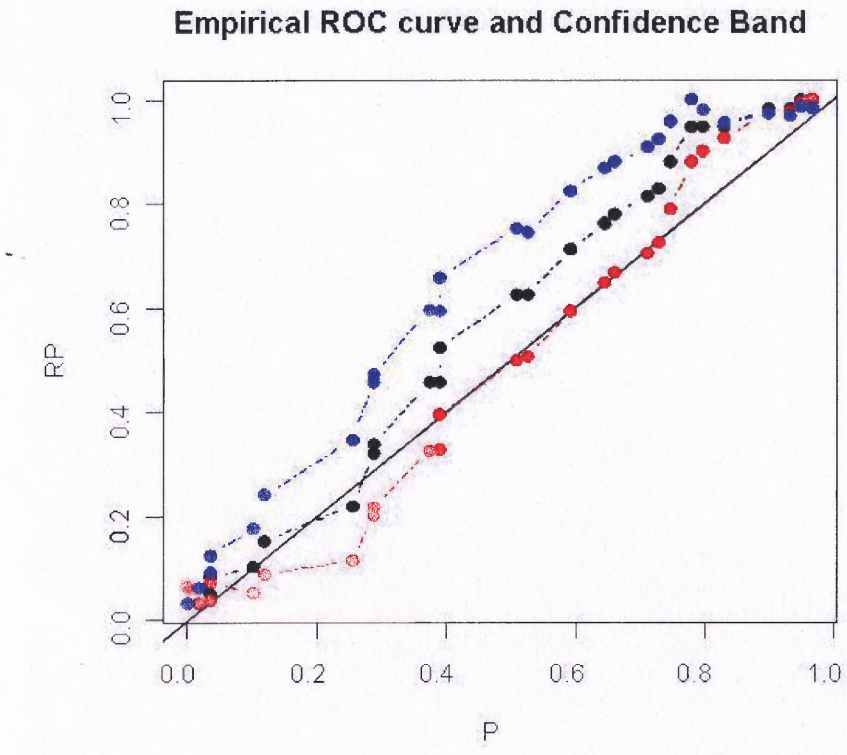

Figure 5.4 Empirical ROC curve and a 95% confidence band.

Based on the method described in the beginning of this chapter (see Figure 5.2), an estimate of the area under the empirical ROC curve is computed. The value of the probability corresponding to the optimal point of separation is given by 0.779661, the notation for this point is p^*. The corresponding ROC curve value $R(p^*)$ is given by 0.9491525. Therefore the estimated area under the ROC curve given by $\left[\frac{R(p^*)-p^*+1}{2}\right]$, is 0.5847458 $[\approx P(Y > X)]$. Given in the following Table 5.1 is the comparison of the estimates from different methods. The same example of the double-blind drugs is used in this effect size measurement.

Table 5.1 Effect Size Measures from Different Methods of Estimation.

WMW	Smoothed*	Empirical(1)**	Empirical(2)†	Empirical(3)°
0.57541	0.56744	0.5847458	0.5754094	0.54869290

* Based on the smoothed density estimate (See Chapter 4).
** Based on optimal point of separation (See earlier this chapter).
† Based on numerical Riemann-Stieltjes integration.
° Based on Empirical distribution function.

Riemann integration deals with finding the area under a curve $f(x)$ in a bounded interval $[a,b]$, where this interval is divided into finite number of subintervals n. Then the sum of the form $\sum_{k=1}^{n} f(t_k)\Delta x_k$ is considered where t_k is a point in the k^{th} interval. This way the area is computed by the means of rectangles.

A more general case of the Riemann integration is the *Riemann-Stieltjes integration* (see Apostol 1990). This involves two functions. In the present discussion, these two functions are $p(c)$ and $R(p(c))$ respectively. This would deal with the sum of the form $\sum_{k=1}^{n} R(t_k)\Delta p_k$ (Recall that the original variable c, the threshold value, is not represented on the ROC curve). To get an estimate of the ES, the mean of the upper sum and the lower sum of the integral is computed using a numerical Riemann-Stieltjes integral.

5.5 Numerical Results

The confidence interval of the ROC curve based on the expression given in the equation (5.2) is given below (Table 5.2). The reported values of p and $R(p)$ are the empirical values computed from the data set given in table 2.1. The confidence interval is the SRS based result simplified from the general result for GRSS, Ghosh and Tiwari (2007).

The Confidence interval of the empirical ROC curve is reasonably sharp. Note that it is actually a point wise confidence band based on the standard deviation of the ROC curve evaluated at a grid of points using the Result 5.4.2.

Table 5.2 Confidence Interval of ROC Curve

p	$\hat{R}(p)$	$\sqrt{\hat{K}(p)}$	Lower	Upper
0.01694915	0.03389831	0.1724763	0.002777945	0.06501867
0.03389831	0.05084746	0.2427279	0.007051432	0.09464348
0.10169492	0.10169492	0.4420965	0.021926284	0.18146355
0.11864407	0.15254237	0.5296653	0.056973464	0.24811128
0.25423729	0.22033898	0.7565856	0.083826255	0.35685171
0.28813559	0.33898305	0.8291645	0.189374721	0.48859138
0.38983051	0.45762712	0.8576754	0.302874501	0.61237974
0.52542373	0.62711864	0.7957604	0.483537499	0.77069979
0.59322034	0.71186441	0.7765049	0.571757598	0.85197122
0.72881356	0.83050847	0.6910311	0.705823921	0.95519303
0.74576271	0.88135593	0.6062952	0.771960499	0.99075137
0.77966102	0.94915254	0.4686075	0.864600462	1.03370462

The estimating algorithm for the optimal threshold applied to the combination of normal variates yields the following results. Samples of size m, where m denotes 200 or 2000, are drawn from the Y and X densities and C* points are estimated from the fitted densities (see section 5.3 in the current chapter). Reported values are the median estimates obtained from 40 replications of the process. Given in parenthesis are the

median absolute deviations of the estimates.

Table 5.3 Estimated Optimal Threshold Value c*

Y	X	C^*	$\widehat{C^*}$(m=2000)	$\widehat{C^*}$(m=200)
$N(1,1)$	$N(0,1)$	0.5	0.503(0.057)	0.397(0.113)
$N(1/2,1)$	$N(0,1)$	0.25	0.237(0.090)	0.191(0.362)
$N(2,1)$	$N(0,1)$	1	0.999(0.045)	0.999(0.113)
$N(2,1)$	$N(1,1)$	1.5	1.473 (0.068)	1.488(0.192)
$N(3/2,1)$	$N(1,1)$	1.25	1.267(0.147)	1.107(0.520)
$N(3,1)$	$N(1,1)$	2	2.007(0.045)	2.007(0.158)

The noteworthy observation here is that the estimates are reasonably stable even when densities are estimated using fewer numbers of data points, although the mean squared errors are relatively higher for the small sample size.

CHAPTER 6

CONCLUSIONS AND FUTURE STUDIES

The major finding of this study is the fact that information from datasets with heavy dependence can be extracted with the use of tools provided. Also, density estimation methods, and particularly kernel smoothing can be used in efficacy measurement. The estimated density has the nice root-n consistency and yield smaller MISE compared to the density estimate computed from independent data. Both have the same form of the bias function.

Confidence intervals based on bootstrap and double bootstrap methods are quite effective in the context of finite or infinite dimensional problems like density estimation.

Given that in the real world one has to work with smaller datasets which are often imperfect and dependent, this study is an example where heavy computing can certainly overcome some of these shortcomings in an effective way. Efficacy measurement is a problem that is constantly under focus in clinical studies. In the current age of advanced computing, both the problem and the approaches towards its solution can be addressed through developing efficient algorithms and programs.

There are many possible ways the current study can be extended in different directions:

(1) Theoretical and numeric implementation and development of double bootstrap in dependent data.

(2) How to find a theoretical estimate of the density when the D data with the variance covariance matrix = $I \otimes P_1 + (J-I) \otimes P_2$ is either known or estimable.

(3) How dependent density estimation can be utilized in non-parametric regression (smoothing) methods in a dependent set up.

(4) Generalize simulation study to include normal mixture densities.

(5) Apply the nonparametric Bayesian method to estimate the ROC curve using Mixture of Dirichlet Process (MDP) priors. This can be done by proceeding as in Erkanli et al. (2006).

(6) Some asymptotic results on the general form of the kernel covariate. For example if the underlying distribution is normal, closed form expressions can be obtained for the covariance after the application of the kernel function. Similarly, assuming known continuous distributions of the underlying densities would give us a comparison of the results.

APPENDIX

SOURCE CODES

Some selected source codes written in R are included in this section. R is the free version of the S/Splus software (Ihaka and Gentleman, 1996). The R packages along with relevant packages are available for download for free from the comprehensive R network, http://cran.r-project.org/.

The R version 2.3.1 is used for the present study along with the editor Tinn-R version 1.17.2.4, available for download from the following website http://www.sciviews.org/Tinn-R/.

Each of these codes is linked with one or more chapters. There are several lines of comments included in each code to clearly explain the objective.

The following program is used to compute the UMVU estimator in the exponential case. See Chapter 3 for the related derivations.

```
# PROGRAM 1

##################################
# PROGRAM TO CLACULATE THE UMVUE #
##################################

myTheta=function(s,t,m,n)
{
if(s<t)
{
  su=0
 nn=n-1
  for (i in 2:nn)
{
k=i-1
  nk=n-k-1
  su=su+exp( nk*log(1-s/t) + k*log(s/t)+ sum(log(1:nn)) - sum(log(1:k))
-
sum(log(1:nk)) )/(m+k-1)
}
su=su*(m-1)
}

else if(s>t)
{
  su=0
 mm=m-2
```

```
  for (i in 2:mm)
{
k=i-1
  mk=mm-k
  su=su+exp( mk*log(1-t/s) + (k+1)*log(t/s)+ sum(log(1:mm)) -
sum(log(1:k)) -
sum(log(1:mk)) )/(n+k-1)
}
su=su*(m-1)
}

return(su)
}
# END OF PROGRAM 1 #
```

The following code is used to do the Monte Carlo simulation of the estimator, by repeatedly calling the above routine. The outputs are recorded in a file and saved for post processing.

```
# PROGRAM 2

#####################################
# MONTE CARLO SIMULATION FOR UMVUE #
#####################################
MM=5000;A=3;B=4;
for (i in 1:MM)
{
#####################################
# Generate a few exponential random
# variables with mean lambda_1 and 2
#####################################
lambda_1=A;lambda_2=B;m=300;n=200;
x=rexp(m,1/lambda_1); y=rexp(n,1/lambda_2);
s=sum(x);t=sum(y);
#########################################
# Initialize the values of the estimates
#########################################
thetaMLE  =    0;
thetaMLE = mean(x)/(mean(x)+mean(y));
#################################
# Our estimate of this Quantity
#################################
####################
# usage of myTheta #
# myTheta(s,t,m,n) #
####################

#cat("real theta=",lambda_1/(lambda_1+lambda_2),"\n")
#cat("myTheta=",1-myTheta(s,t,m,n),"\n","thetaMLE=", thetaMLE,"\n")
# file = "c:/courses/research/clinical/exp_UMVU.txt"

ss = c(lambda_1/(lambda_1+lambda_2),thetaMLE, 1-myTheta(s,t,m,n));
```

```
write(ss,file = "c:/DATA/exp_UMVU_A_by_B.txt",
      ncolumns=length(ss),sep="\t", append=T);
rm(ss);
}
# END OF PROGRAM 2 #
```

The following program calculates the bootstrap density estimator from the dataset introduced in Chapter 2 and used repeatedly to illustrate many of the methods that is developed in the current study.

```
# PROGRAM 3

#######################################################
# CREATE BOOTSTRAP DENSITY ESTIMATE FROM DRUG DATA #
#######################################################
# Read the original data #
bp=read.csv("L:/courses/research/clinical/data_bp_a_b.csv",
header=TRUE)
# Calculate the difference #
D = Diff(bp[,1],bp[,2])
# Prepare for Bootstrap #
B = 5000
ss=0;

#######################################################
# Create the File names temp_1.txt to temp_5000.txt
#######################################################
File  = paste("D:/DATA/temp_",1:B,".txt", sep="")
# File  = paste("C:/DATA/temp_",1:B,".txt", sep="") # For school
for (i in 1:B)
{
DD = sample(D, replace=T)
bw = bw.nrd0(DD)
ss = density(DD, bw="nrd0",kernel="epanechnikov", n=512,from=-80,to
=80)
# Density has been evaluated at the fixed x values #
ss=cbind(ss$x,ss$y) # Save just the x and y's
write.table(ss, file=File[i]) # Use write table to write in the files
}
dd= mat.or.vec(512,B) # Declare the place holding variable dd
for (i in 1:B)
{
tt= read.table(File[i]) # Read the tables in this temporary var tt
dd[,i]=tt$V2            # Save only the evaluated density
}

M = 300
x           = rnorm(round(M/2),0,1);
y           = rnorm(round(M/2),1,1);
z           = rnorm(M,1,2);
Dxy = Diff(x,y);
dxy   = density(Dxy, bw="bcv", kernel="epanechnikov", from=-8,to =8,
n=1024);
```

```
dz     = density(z, bw="bcv", kernel="epanechnikov", from=-8,to =8,
n=1024);
Z      = dnorm(dxy$x, mean=1, sd=sqrt(2), log = FALSE)

part=seq(1,1024,by=20);
plot(Z[part],cex=.75,lty=1,type="b",xaxt="n",ylab="density", xlab="x
values",main="density estimates compared with the real density",
ylim=c(0,.4), col="black", pch=4,lwd=2)

# cbind(part,dxy$x[part])
axis(1,at=c(1,10,20,30,39,49), labels=c(-8,-5,-2,1,4,7), las=1)
points(dz$y[part], col="red", pch=15, cex=.75,lty=2, type="b",lwd=2)
points(dxy$y[part], col="blue", pch=24 , cex=.75,lty=3, type="b",lwd=2)
leg.txt=c("real density", "dens-est indep", "dens-est dep");
legend(0.3, leg.txt, col = c("black","red","blue"),
pch=c(4,15,24),lty=c(1,2,3),
lwd=c(2,2,2))


# Read the original data #
bp=read.csv("L:/courses/research/clinical/data_bp_a_b.csv",
header=TRUE)
# Calculate the difference #
D = Diff(bp[,1],bp[,2])
B = 5000
ss=0;

dd= mat.or.vec(512,(B+1)) # Declare the place holding variable dd
# Fill the last column with the x values
dd[,(B+1)]=density(D, bw="nrd0",kernel="epanechnikov", n=512,from=-
80,to =80)$x

# Now start the bootstrap resampling

for (i in 1:B)
{
DD = sample(D, replace=T)
ss = density(DD, bw="nrd0",kernel="epanechnikov", n=512,from=-80,to
=80)
# Density has been evaluated at the fixed x values #
dd[,i]=ss$y # Save just the y
}

V = seq(-40,40,by=5)
DD= mat.or.vec(17,B)
for(i in 1:B)
{ DD[,i]=apply(as.array(V),1,PyLTC,dd[,B+1],dd[,i])}
M      = apply(as.array(DD),1,mean)
# SS     = apply(as.array(DD),1,var)
MSE    = apply((DD-M)^2,1,mean)
# The probabilities of the values given with the MSE's
cbind(V, round(1-M,digits=5), round(MSE,digits=8))

# Bootstrap 95% CI of the estimated probabilities.
apply(as.array(DD),1,quantile, prob=c(0.025,0.975))

# END OF PROGRAM 3 #
```

The following program is self explanatory. The ultimate goal of this program is to utilize the density estimate and find the efficacy measure following the methods described in Chapter 4.

```
# PROGRAM 4

#####################################################
#                                                   #
# This program would do the following things:       #
#                                                   #
# (1) Draw B bootstrap samples                      #
#                                                   #
# (2) Compute the densities from them               #
#                                                   #
# (3) Save the pointwise median (estimate)          #
#                                                   #
# (4) Save 97.5th and 2.5th percentile (CI)         #
#                                                   #
# (5) Calculate P(c1<D<c2) for a range of c's       #
#                                                   #
# (6)(1)Draw B btstrp smpl from N(1,1)\ N(0,1)      #
#                                                   #
# (7)(2) compute densities                          #
#                                                   #
# (8)(3) save the pointwise median                  #
#                                                   #
# (9)(4) Save 97.5th and 2.5th percentile (CI)      #
#                                                   #
#(10)(5) Calculate P(c1<D<c2) for a range of c's    #
#                                                   #
#(11)*** Compare the findings with real values      #
#                                                   #
#(12)compute percentiles from the density           #
#                                                   #
#(13)*** TRUE/EST/CI-density on the same plot!!!    #
#                                                   #
#(14) See what percent of the points are captured   #
#                                                   #
#(15) Do the same experiment 50 times and compute   #
#     boot mean and variance                        #
#####################################################


bp=read.csv("F:/courses/research/clinical/data_bp_a_b.csv",
header=TRUE) #NJIT
#bp=read.csv("L:/courses/research/clinical/data_bp_a_b.csv",
header=TRUE) # home
xy = Diff(bp[,1],bp[,2])

lowxy      = min(xy) - round(0.25*sd(xy))
hixy       = max(xy) + round(0.25*sd(xy))
B=20;
nn=512
# nn =512
```

```
# File  = paste("D:/DATA/PyLTC/BOOTdensity_",1:B,".txt", sep="") # home
File  = paste("C:/DATA/BOOTdensity_",1:B,".txt", sep="") # NJIT

for (i in 1:B)
     {
     XY = sample(xy, replace=T)
     dXY = density(XY, bw="bcv", kernel="epanechnikov",from=lowxy,
to=hixy,n=nn)
     dXY=cbind(dXY$x,dXY$y) # Save just the x and y's
     write.table(dXY, file=File[i]) # Use write table to write in the
files
     }

# READ IN NOW

dd= mat.or.vec(512,B) # Declare the place holding variable dd

for (i in 1:B)
      {
      tt= read.table(File[i]) # Read the tables in this temp var tt
      dd[,i]=tt$V2            # Save only the evaluated density
      }

      rm(tt)# Now just delete tt and retain dd.

# This is our density it would have
# the x values, median values/ 2.5%
# and 97.5% values &var; 5 columns


our_den     = mat.or.vec(512,5)
our_den[,1] = dXY[,1]
our_den[,2]     = apply(dd,1,median);
our_den[,5]     = apply(dd,1,var);
# our_den[,c(2,5)]      =  cbind(apply(dd,1,median),apply(a,1,var))
for (i in 1:512)
      {
      our_den[i,3] = quantile(dd[i,],prob=.025);
      our_den[i,4] = quantile(dd[i,],prob=0.975);
      }

# par(cex=0.7, cex.axis=1.2, cex.lab=1.2, cex.main=1.2)
plot(our_den[,2], pch='+', ylim=c(0,.035))
points(our_den[,3], col="red", pch="^")
points(our_den[,4], col="blue", pch="*")


#points(our_den[,2]+20000*our_den[,5], col="grey80", pch="@", type="b")
#points(our_den[,2]-20000*our_den[,5], col="khaki", pch="%", type="b")
CC = seq(-20, 60, by=5)
PyLTC(CC[1:3],our_den[,1],dd[,1])
sapply(t(CC),PyLTC,X = our_den[,1],Y=dd[,1])

myfun <- function(X,y) {y[which.min(abs(X-y))]}
apply(as.array(X),1,myfun, y = our_den[,1])
# rm(B,bp,D,dXY,File,dd,our_den)
```

 # END OF PROGRAM 4 #

The following program will be used to compute the covariance kernel of the Gaussian process under a specific point P. The relevant theory can be found in Chapter 5. Note that this is a three step transition, once from GRSS/BRSS to SRS, then from the shift function to the ROC, then both the evaluation points for the kernel is taken to be the same. In other words, $K_{R,SRS}(p, q)$ is taken to be $K_{R,SRS}(p, p)$. See section 5.4 in Chapter 5 for more information.

```
# PROGRAM 5

##############################################
# This function would compute the covariance #
# kernel of a given pair of points, of the   #
# vertical shift function                    #
# This function would later be used for      #
# a nonparametric CI of the ROC curve!!      #
##############################################

# There would be three functions used as subfunctions
# (a) density
# (b) Finv /cdfinv
# (c) G.Finv

covKern = function(P, X, Y)
{
P=1-P
del = (dense(Finv(P,X),Y,8))^2/((dense(Finv(P,X),X,8))^2)*(1/0.5)*(P-
P^2)
+     (1/0.5)*(GFinv(P,X,Y)*(1-GFinv(P,X,Y)))
#if(p==0 || p==1) del=0;
return(del)
}

dense = function(x, xdata,bw)
{
n           = length(xdata)
xdata      = sort(xdata)
idx         = which(abs(xdata-x)<bw)
return(3/(4*n*bw)*sum(1-((xdata[idx]-x)/bw)^2))
}

Finv = function(p,anydata)
{
n=length(anydata)
anydata = sort(anydata)
return(anydata[ceiling(n*p)])
}

GFinv = function(p,xdata,ydata)
{
m=length(xdata)
```

```
n=length(ydata)
xdata=sort(xdata)
ydata=sort(ydata)
(1/n)*sum(ydata<xdata[ceiling(m*p)])
}

# as.numeric(unlist(lapply(seq(.01,.99,by=.01),covKern,bp[,1],bp[,2])))

# END OF PROGRAM 5 #
```

The following function will compute the point wise confidence interval for the empirical ROC curve based on the covariance kernel explained above. To change the coverage probability only the value of the quantile function from the standard normal distribution needs to be changed. These results just like the results used in the above program are given in section 5.4 of Chapter 5.

```
# PROGRAM 6

#####################################################################
# This program will create the CI based on the general GRSS       #
# formula simplified for the SRS case, it would also plot the     #
# CI with the empirical ROC function as well as the smoothed      #
# version of the ROC curve that would be built on the basis       #
# of the nonparametric regression (Nadaraya-Watson) curve.        #
#####################################################################

#bp=read.csv("F:/courses/research/clinical/data_bp_a_b.csv",
header=TRUE) #NJIT
#bp=read.csv("L:/courses/research/clinical/data_bp_a_b.csv",
header=TRUE) # home
xy = Diff(bp[,1],bp[,2])
lowxy       = min(xy) - round(0.25*sd(xy))
hixy        = max(xy) + round(0.25*sd(xy))

V=density(bp[,1], bw="bcv", kernel="epanechnikov", from = lowbp, to
=hibp)
W=density(bp[,2], bw="bcv", kernel="epanechnikov", from = lowbp, to
=hibp)

FINE=100;
X = seq(lowbp, hibp, by = (hibp-lowbp)/FINE);
X = rev(X);

comp <- function(X,LL) {sum(LL>X)}
P     =     apply(as.array(X),1, comp, LL=bp[,1])/59;
RP    =     apply(as.array(X),1, comp, LL=bp[,2])/59;

plot(RP~P, type="b", lty = 4, pch=16, main="Empirical ROC curve of
drugs A and B")
abline(0,1)
```

```
##############################
# USE the covKernNew routine  #
# for several values of  P    #
##############################

P      = P[41:100] # choose a smaller set of P
RP     = RP[41:100] # choose a smaller set of RP as well

# K    =lapply(as.array(P),FUN=covKernNew,V=V, W=W)
# K    =lapply(as.array(P),covKern,V$x, W$y)

K    = as.numeric(unlist(K))
K    = sqrt(K)

cbind(RP-(1.96*K/sqrt(118)),RP+(1.96*K/sqrt(118)))

cbind(P,RP,K,(1.96*K/sqrt(118)));
```

 # END OF PROGRAM 6 #

REFERENCES

Acion L., Peterson J. J., Temple S. and Arndt S. (2006). Probabilistic Index: an Intuitive Non-parametric Approach to Measuring the Size of Treatment Effects. *Statistics in Medicine* **25** 4, 591-602.

Agresti A. (2002). *Categorical Data Analysis.* Wiley Series in Probability and Statistics.

Ali M M., Pal M, and Woo J. (2005). Inference on P(Y < X) in Generalized Uniform Distributions. *Calcutta Statistical Association Bulletin* **57**, 35-48.

Ali M M.,Woo J. and Pal M. (2004). Inference on Reliability P(Y < X) in a Two-parameter Exponential Distribution. *International Journal of Statistical Sciences* **3**, 119-125.

Apostol T.M. (1990). *Mathematical Analysis, Second Edition*, Addison-Wesley.

Azzalini A. (1981). A Note on the Estimation of a Distribution Function and Quantiles by a Kernel Method. *Biometrika* **68** 1 326-328.

Bergmann R., Ludbrook J. and Spooren P. J. M. W. (2000). Different Outcomes of the Wilcoxon-Mann-Whitney Test from Different Statistics Packages. *The American Statistician* **54** 1,72-77.

Bernhard G., Alle M., Herbold M., and Meyers W. (1988). Investigation on the Reliability of Some Elementary Nonparametric Methods in Statistical Analysis Systems. *Statistical Software Newsletter* **4**, 19-26.

Birnbaum Z. W., McCarty R. C. (1958). A Distribution-Free Upper Confidence Bound for Pr{Y < X}, Based on Independent Samples of X and Y. *The Annals of Mathematical Statistics* **29** 2, 558-562.

Bowman A.W. and Azzalini A. (1997). *Applied Smoothing Techniques for Data Analysis, The Kernel Approach with S-Plus Illustrations.* Oxford Statistical Science Series. Oxford University Press.

Brumback L. C, Pepe M.S. and Alonzo T. A. (2006). Using the ROC Curve for Gauging Treatment Effect in Clinical Trials. *Statistics in Medicine* **25** 4, 575-590.

Chen M. and Kianifard F. (2000). A nonparametric procedure associated with a clinically meaningful efficacy measure. *Biostatistics* **1**, 293-298.

Chen S. X. (1996). Empirical Likelihood Confidence Intervals for Nonparametric Density Estimation. *Biometrika* **83** 2, 329-341.

Cramer E. and Kamps U. (1997). A Note on UMVUE of Pr(X < Y) in the Exponential Case. *Communications in Statistics – Theory and Methods* **26**, 1051-1055.

Doksum K. A. (1974). Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case. *The Annals of Statistics* **2**, 267-277.

Dodd L. E. and Pepe M.S. (2003). Semiparametric Regression for the Area Under the Receiver Operating Characteristic Curve. *Journal of the American Statistical Association* **98**, 409-417.

Efron and Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman & Hall/CRC.

Erkanli A., Minje S., Costello E. J. and Adrian A. (2006). Bayesian Semi-parametric ROC Analysis. *Statistics in Medicine* **25**, 3905-3928.

Faraggi D. and Reiser B. (2002). Estimation of the Area Under the ROC Curve. *Statistics in Medicine* **21** 20, 3093-3106.

Ghosh K. and Tiwari R.C. (2007). Empirical Process Approach to Some Two Sample Problems Based on Ranked Set Samples. *Ann Inst Stat Math* (in press).

Hall, P. G. and Hyndman, R. J. (2003). Improved Methods for Bandwidth Selection When Estimating ROC Curves. *Statistics & Probability Letters* **64** 2, 181-189.

Hall P.G., Lahiri S.N. and Truong Y. K. (1995). On Bandwidth Choice for Density Estimation With Dependent Data. *The Annals of Statistics* **23** 6, 2241-2263.

Hanley J.A. and McNeil B.J. (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **143**, 29-36.

Hardle W., Muller M., Sperlich S. and Werwatz A. (2004). Nonparametric and Semiparametric Models. *Springer Series in Statistics.*

Hogg R.V., Ruberg S.J. and Yuh L. (1990). *Robust Data Analysis. In Statistical Methodology in the Pharmaceutical Sciences*, ed. D.A. Berry, New York: Mercel-Dekker. 357-388.

Ihaka R. and Gentleman R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5** 3, 299-314.

Ivshin V.V. (1996). Unbiased Estimators of Pr (X < Y) and Their Variances in the Case of Uniform and Two-parameter Exponential Distributions. *Journal of Mathematical Science* **81**, 2790-2793.

Izenman A. J. (1988). Reviewed Work(s): Density Estimation for Statistics and Data Analysis. by B. W. Silverman. *Journal of the American Statistical Association* **83** 401, 269-270.

Izenman A. J. (1991). Recent Developments in Nonparametric Density Estimation. *Journal of the American Statistical Association* **86** 413, 205-224.

Ker A. P. and Ergün A.T. (2005). Empirical Bayes Nonparametric Kernel Density Estimation. *Statistics & Probability Letters* **75** 4, 315-324.

Lawson A. B. and Williams F. L. R. (1994). Armadale: A Case-Study in Environmental Epidemiology. *Journal of the Royal Statistical Society. Series A* **157** 2, 285-298.

Lee M. T., Rosner B. A. (2001). The Average Area Under Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach Based on Generalized Two-Sample Wilcoxon Statistics. *Applied Statistics* **50** 3, 337-344.

Lehman E.L.(1963). Robust Estimation in Analysis of Variance. *The Annals of Mathematical Statistics* **34** 3, 957-966.

Lehmann E. L. (1998). *Nonparanletrics: Statistical Methods Based on Ranks (revised 1st ed.),* Upper Saddle River, NJ: Prentice Hall.

Lloyd C. J. (2002). Estimation of a Convex ROC curve. *Statistics & Probability Letters* **59** 1, 99-111.

Lloyd C. J. and Zhou Y. (1999). Kernel Estimators of the ROC Curve are Better than Empirical. *Statistics & Probability Letters* **44** 3, 221-228.

Lloyd C. J. (1998). Using Smoothed Receiver Operating Characteristic Curves to Summarize and Compare Diagnostic Systems. *Journal of the American Statistical Association* **93** 444, 1356-1364.

Ludbrook, J. (1995). Microcomputer Statistics Packages for Biomedical Scientists. *Clinical and Experimental Pharmacology and Physiology* **22**, 976-986.

Mann H. B. and Whitney D. R.(1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18** 1, 50-60.

Miller, K. S. (1981) On the Inverse of the Sum of Matrices (in Notes). Mathematics Magazine **54** 2, 67-72.

McCool J.I.(1991). Inference on P(X < Y) in the Weibull Case. *Communications in Statistics – Simulation and Computation* **20** 1, 129-148.

Newcombe R. G. (2006 a). Confidence Intervals for an Effect Size Measure Based on the Mann-Whitney Statistic. Part 1: General Issues and Tail-area-based Methods. *Statistics in Medicine* **25** 4, 543-557.

Newcombe R. G. (2006 b). Confidence Intervals for an Effect Size Measure Based on the Mann-Whitney Statistic. Part 2: Asymptotic Methods and Evaluation. *Statistics in Medicine* **25** 4, 559-573.

Owen D. B., Craswell K. J. and Hanson D. L. (1964). Nonparametric Upper Confidence Bounds for Pr{Y < X} and Confidence Limits for Pr{Y < X} When X and Y are Normal. *Journal of the American Statistical Association* **59** 307, 906-924.

Pal M., Ali M M. and Woo J. (2005). Estimation and Testing of P(Y>X) in Two-parameter Exponential Distributions. *Statistics* **39** 5, 415-428.

Pepe M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford:Oxford University Press.

Pepe M.S. (1997). A Regression Modeling Framework for Receiver Operating Characteristic Curves in Medical Diagnostic Testing. *Biometrika* **84** 3, 595-608.

Potthoff R. F. (1963). Use of the Wilcoxon Statistic for a Generalized Behrens-Fisher Problem. *The Annals of Mathematical Statistics* **34** 4, 1596-1599.

Puri P.S. and Rubin H (1970). A Characterization Based on the Absolute Difference of Two i.i.d. Random Variables. *Annals of Mathematical Statistics* **41**, 2113-2122.

Rosenblatt M. (1956). Remark on Some Nonparametric Estimates of a Density Function. *Annals of Mathematical Statistics* **27**, 832-837.

Sathe Y. S. and Shah S. P. (1981). Estimating P(X > Y) for the Exponential Distribution. *Commun. Stat. Theory Meth.* **A10** 1, 39-47.

Saw J. G. (1966). A Non-Parametric Comparison of Two Samples One of Which is Censored. *Biometrika* **53** 3-4, 599-602.

Scott D. W. (1992). Multivariate Density Estimation: Theory, Practice, and Visualization. *Wiley-Interscience.*

Seaman D. E. and Powell R. A. (1996). An Evaluation of the Accuracy of Kernel Density Estimators for Home Range Analysis. *Ecology* **77** 7, 2075-2085.

Seheult A. H and Quesenberry C. P. (1971). On Unbiased Estimation of Density Functions. *The Annals of Mathematical Statistics* **42** 4, 1434-1438.

Silverman, B W (1986). *Density Estimation for Statistics and Data Analysis.* London:Chapman & Hall.

Simonoff J. S., Hochberg Y. and Reiser B. (1986). Alternative Estimation Procedures for Pr(X < Y) in Categorized Data. *Biometrics* **42** 4 , 895-907.

Taylor C. C. (1989). Bootstrap Choice of the Smoothing Parameter in Kernel Density Estimation. *Biometrika* **76** 4, 705-712.

Terrell G.R. and Scott D.W. (1992). Variable Kernel Density Estimation. *Annals of Statistics* **20** 3, 1236-1265.

Titterington D. M. and Mill, G. M. (1983). Kernel-Based Density Estimates from Incomplete Data. *Journal of the Royal Statistical Society. Series B* **45** 2, 258-266.

Van Belle G. (2002). *Statistical Rules of Thumb.* Wiley Series in Probability and Statistics.

Venkatraman E. S. and Begg, C. B. (1999). Properties of a Nonparametric Test for Early Comparison of Treatments in Clinical Trials in the Presence of Surrogate Endpoints. *Biometrics* **55** 4, 1171-1176.

Walter S. D. (2002). Properties of the Summary Receiver Operating Characteristic (SROC) Curve for Diagnostic Test Data. *Statistics in Medicine* **21** 9, 1237-1256.

Wilcoxon F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1** 6, 80-83.

Wolfe D. A. and Hogg R.V. (1971). On Constructing Statistics and Reporting Data *The American Statistician* **25** 4, 27-30.

Zhou X. and Harezlak J. (2002). Comparison of Bandwidth Selection Methods for Kernel Smoothing of ROC Curves. *Statistics in Medicine* **21** 14, 2045-2055.