1997

A nonparametric approach to estimating the overlapping coefficient using the kernel estimation technique.

Traci Elnora Clemons
University of Alabama at Birmingham

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

A NONPARAMETRIC APPROACH TO ESTIMATING THE OVERLAPPING COEFFICIENT USING THE KERNEL ESTIMATION TECHNIQUE

by

TRACI E. CLEMONS

A DISSERTATION

Submitted to the graduate faculty of the University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

1997

UMI Number: 9734423

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

ABSTRACT OF DISSERTATION
GRADUATE SCHOOL, UNIVERSITY OF ALABAMA AT BIRMINGHAM

Degree Ph.D. Program Biostatistics

Name of Candidate Traci E. Clemons

Committee Chair Dr. Edwin Bradley, Jr.

Title A Nonparametric Approach to Estimating the Overlapping Coefficient Using the Kernel Estimation Technique

This study examines the sampling behavior of the overlapping coefficient, OVL. The OVL is a proposed measure of the agreement between two probability distributions. The OVL is defined for the continuous case as

$$OVL = \int_x \min\,[f_1(x)\,,f_2(x)]\,dx\;;$$

where $f_1(x)$ and $f_2(x)$ are the probability density functions for two distributions of interest. In addition, OVL = 1 - D, where D is the usual index of dissimilarity, but defined for continuous as well as discrete distributions.

Here the properties and sampling behavior of a nonparametric estimator of the OVL are investigated. The nonparametric density estimator chosen to explore the behavior of the OVL is the naive (Rosenblatt) kernel density estimator.

Using Monte Carlo techniques, it is discovered that the sampling estimator of the OVL using the kernel density estimator is biased. The bias of the kernel estimator is a function of the value of the overlap. Also, the bias increases as the similarity of the distributions from which the samples are obtained increases. A bootstrap estimator of the sampling variance of the estimator of the OVL is shown to perform well. The behavior of the sampling estimator of the OVL suggests that the OVL can best serve as a valuable check in investigating the meaningfulness of differences detected between two distributions by other statistical techniques.

ACKNOWLEDGEMENTS

It has been a great pleasure working with Dr. Edwin Bradley, my advisor. I am grateful to the members of my committee, Drs. Charles Katholi, Tonya Smoot, Mary Hovinga, and Pauline Jolly. I am particular indebted to Dr. Charles Katholi for his help in developing the fortran programs used for this study. I wish to express my sincere appreciation to the entire Biostatistics department.

Finally, I acknowledge, with thanks, my indebtedness to my family for their patience and support throughout the time this project was in progress and all through my academic endeavors.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

I. INTRODUCTION

Consider two probability distributions with densities denoted $f_1(x,\lambda_1)$ and $f_2(x,\lambda_2)$ respectively, where λ_1 and λ_2 are parameters and the distributions are of the same parametric form. The two distributions may be said to differ if $\lambda_1 \neq \lambda_2$. The two parameters may in fact differ, yet be somewhat similar in magnitude, thus implying that the two probability distributions may be similar. It may also be the case that we have two distributions of different parametric form, denoted $f_1(x,\lambda_1)$ and $f_2(x,\xi_2)$, which cannot be identical, yet the parameters λ_1 and ξ_2 may be similar.

In more practical cases, we may have two samples from two distributions. Assuming the forms of the distributions to be $f_1(x,\theta_1)$ and $f_2(x,\theta_2)$, where $\theta_1 \neq \theta_2$, the two distributions can be shown to differ using an appropriate statistical test for the equality of the parameters θ_1 and θ_2. Since the power of statistical tests is related to both the magnitude of the difference of the parameters and the sample size from which the parameters are estimated, small differences in the parameters can be declared statistically significant while the true similarity of the two populations of interest goes undetected.

This study explores a measure of agreement between two probability distributions first proposed by Bradley and Piantadosi (1982). The measurement, the overlapping coefficient, denoted OVL, estimates the common area below two probability distributions. The two distributions of interest may be of the same parameter family or from different parametric families. Figure 1 shows the OVL for two normal distributions, while Figure 2 shows the OVL for two Gamma distributions. Bradley and Piantadosi derived the OVL for several cases involving known distributions.

In this work the properties and sampling behavior of such an estimator of OVL is investigated when sampling from two distributions estimated nonparametrically using the naive kernel density estimator. Also, the properties of a nonparametric bootstrap variance estimator of the overlapping coefficient is explored. The remainder of chaper I contains an historical literature overview as an introduction of the overlapping coefficient. Chapter II provides the development of the kernel estimator and the bootstrap variance estimator of the overlapping coefficient. The Monte Carlo simulation study and its results are also summarized. Chapter III explores an alternative reference rule for the kernel estimator. Chapter IV contains application of the nonparametric estimator of the overlapping coefficient to real sets of data. Lastly, chapter V is a discussion of the research and suggestions for further research.

The Overlapping Coefficient

If we let $f_1(x)$ and $f_2(x)$ be two continuous probability functions defined on a common domain of x, the formal definition of the OVL for the continuous case is

$$OVL = \int_x \min[f_1(x), f_2(x)]\, dx \,. \tag{1}$$

If the two densities of interest are discrete, the definition of the OVL becomes

$$OVL = \sum_x \min[f_1(x), f_2(x)] \,. \tag{2}$$

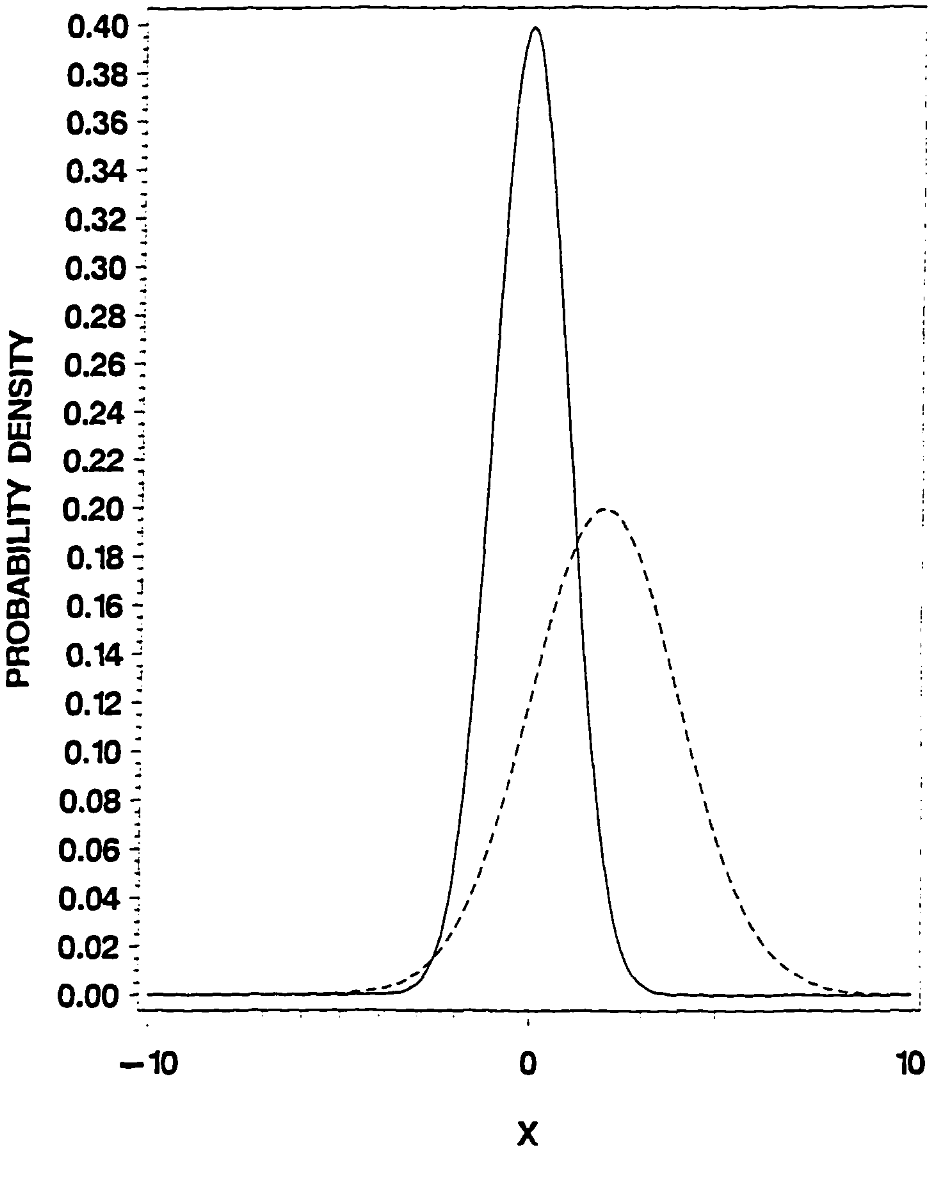

Figure 1. The overlap between two normal distributions. The solid line denotes a standard normal distribution. The dotted line denotes a normal distribution with mean = 2 and variance = 4.

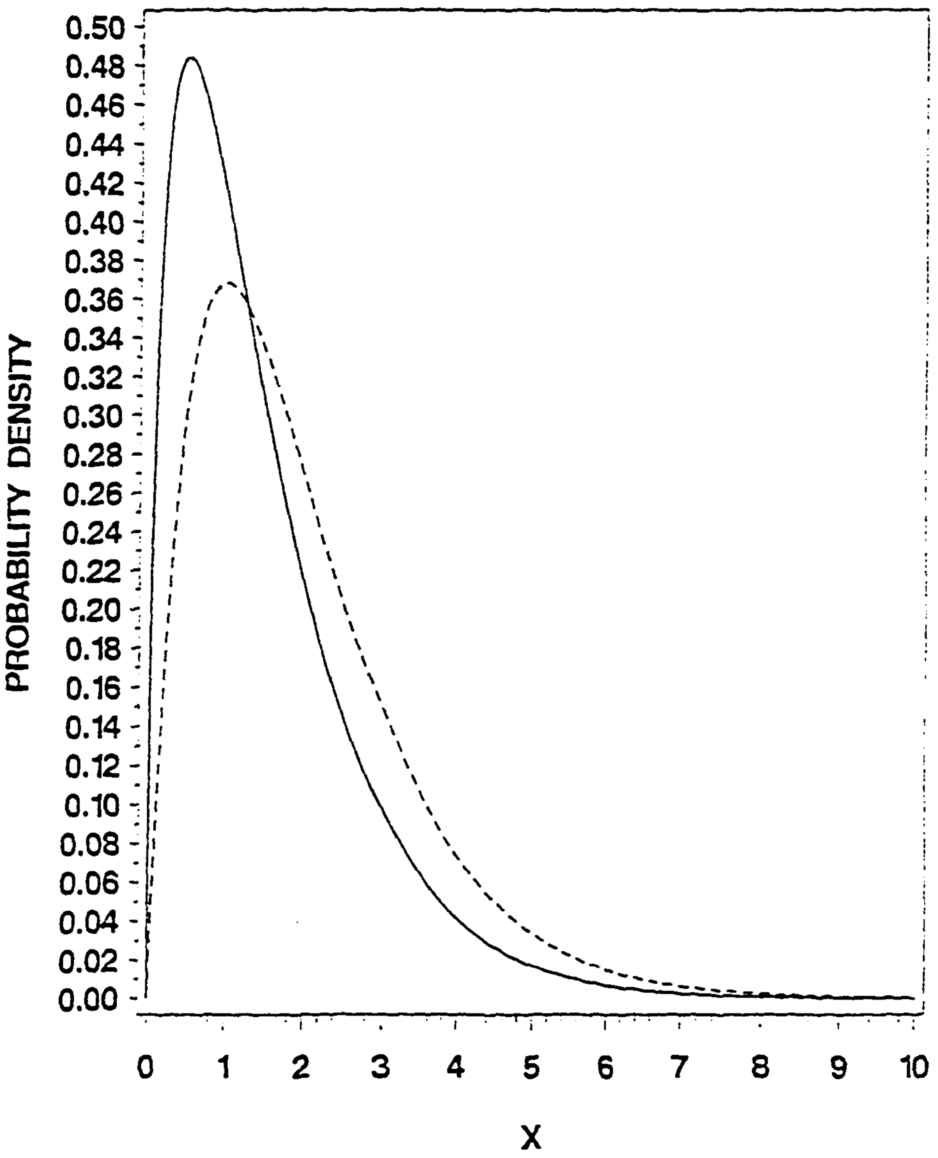

Figure 2. The overlap between two gamma distributions. The solid line denotes a gamma distribution with α=1.5. The dotted line denotes a gamma distribution with α = 2.0.

Bradley and Piantadosi (1982) showed that the OVL has properties that are desirable for any measure of association. First, the OVL ranges between zero and unity. Second, the OVL is unity if and only if the two distributions of interest are identical. Finally, the OVL is zero if and only if the two distributions of interest are completely distinct.

Invariance Property of the OVL

A useful property of the OVL is the invariance property. If we let g(x) be a continuous and differential function defined for all x, then the OVL may be written in terms of this function as follows:

$$OVL = \int_{g(x)} \min[f_1(g(x)), f_2(g(x))]\, dx \ . \qquad (3)$$

The invariance property of the OVL allows for the generalization of estimates of the OVL under normal theory when using normalizing transformations (Tukey, 1975; Box & Cox, 1964).

The Relationship to the Index of Dissimilarity

The OVL is related to what has been known in the literature as the index of dissimilarity, denoted D, which has been commonly used in its discrete form in the context of 2 x C contingency tables. If we use the fact that the two probability density functions are non-negative, the relationship between these two measures can be seen as follows:

$$\min[f_1(x), f_2(x)] = \frac{1}{2}[f_1(x) + f_2(x) - |f_1(x) - f_2(x)|] \ . \qquad (4)$$

Replacing this expression into Equations 1 or 2, it can be seen that OVL = 1 - D; there D in the continuous case is defined as

$$D = \frac{1}{2}\int_x |f_1(x) - f_2(x)|\, dx \qquad (5)$$

and in the discrete case as

$$D = \frac{1}{2}\sum_x |f_1(x) - f_2(x)| \,. \tag{6}$$

Thus the index of dissimilarity is defined as the fraction of probability mass under either distribution not shared with the other. The properties of the OVL apply to D, except that D is zero when the two distributions of interest are identical and is unity when they are completely distinct.

The dissimilarity index dates back to work performed by Karl Pearson in the 1890s. He used a statistic equivalent to 2D as a measure of goodness-of-fit of sample data to some theoretical distribution (Pearson, 1895). Goodman and Kruskal (1979) used D as a measure of association in the context of 2 x C cross classification tables. In other literature, D has been used as an indicator of racial segregation. It was used to compare the relative frequency distribution of African-American and white residents in subdivisions of geographic units (Cortese, Falk, & Cohen, 1976; Duncan & Duncan, 1955). Inman and Bradley (1991) re-examined the behavior of the dissimilarity index under a random allocation model and used it to compare the levels of racial segregation in Birmingham, Alabama and Richmond, Virginia in 1970 and 1980, respectively. They also derived simple approximations for both the mean and variance of D based on a multivariate normal approximation.

Calculation of the OVL Between Known Distributions

Bradley and Piantadosi (1982) present as examples the overlap between two normal distributions, the overlap between the normal and the logistic distributions, and the overlap between two two-parameter exponential distributions. In addition, Inman (1984) presented the overlap between the standard normal and standard Cauchy distributions and two Poisson distributions. In this research, three additional examples are presented.

To determine the OVL between two distributions, it is necessary to determine the point(s) of intersection between the two distributions of interest. The point(s) are found by setting the two distributions equal and solving for the roots of the equation. The resulting equations are non-linear in form. The approach used to find the roots of the non-linear equation in this study is based on the Newton Raphson procedure also refered to as Newton's Method (Hamming, 1971).

Newton-Raphson Method

The Newton-Raphson Method (Newton's Method) is one of the most powerful numerical methods for solving a root-finding problem, Hamming, (1971). It is based on a quadratic Taylor Series expansion. Given a function f(x) which is continuous and twice differentiable on the interval [a,b], let $x_0 \in$ [a,b] be an approximation to the root of the equation, p, such that the first derivative of the function at x_0 differs from zero and $|x_0 - p|$ is small. We consider the first degree Taylor polynomial for the function expanded about x_0, such that:

$$f(x) = f(x_o) + (x - x_o)f'(x_o) + \frac{(x \quad x_o)}{2}f''(\eta(x)), \tag{7}$$

where $\eta(x)$ lies between x and x_o. Since f(p) = 0, equation 7 with $x = p$ is as follows:

$$0 = f(x_o) + (p - x_o)f'(x_o) + \frac{(p \quad x_o)}{2}f''(\eta(p)). \tag{8}$$

If we assume that the term containing $(p-x_0)^2$ is negligible, then solving the above equation for p we obtain the following:

$$p \approx x_o - \frac{f(x_o)}{f'(x_o)}. \tag{9}$$

The Newton's Method involves generating a sequence $\{p_n\}$ defined by:

$$p_n \approx p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad n \geq 1 . \qquad (10)$$

This process is an iterative process which is begun by giving an initial approximation, p_0, which is near the root p. This procedure is repeated until the iteration ultimately coverges to a local, relative maximum, if not to a unique global maximum.

OVL Between Two Gamma Distributions

The density of the Gamma random variable is

$$f(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} \qquad x>0 \quad \alpha , \beta > 0 . \qquad (11)$$

The two densities will intersect at one or two points depending on the shape and scale parameters used. These points can be found by setting the two densities equal and using Newton's method to find the points of intersection. Once the point(s) of intersection are found, equation 1 can be used to compute the value of the OVL.

For example, if $\alpha_1 = 1.5$, $\beta_1 = 1.0$, $\alpha_2 = 2.0$ and $\beta_2 = 1.0$, then the OVL can be computed as follows. As shown previously in Figure 2, the two densities intersect at one point. This point was found by setting the two densities equal and using Newton's method, described previously, to determine the point of intersection, $x_o = 1.274$. Next we use Equation 1 to find the value of the OVL as follows:

$$OVL = \int_0^{1.274} \frac{x_o e^{-x_o}}{\Gamma(2)} + \int_{1.274}^{\infty} \frac{x_o^{0.5} e^{-x_o}}{\Gamma(1.5)} = 0.363938 + 0.466679 = 0.830617 . \qquad (12)$$

OVL Between Two Weibull Distributions

Suppose we have two Weibull distributions with probability density functions defined as follows:

$$f(x) = \alpha_i \beta_j x^{\beta_j -1} e^{-\alpha_i x^{\beta_j}} \qquad x > 0 \qquad \alpha_i , \beta_j > 0 \ . \tag{13}$$

By equating the two densities, we find either one or two points of intersection depending on the values of the shape and scale parameters. Again as with the gamma distribution, the points of intersection can be found by using Newton's method. Once the points are found, Equation 1 can be used to evaluate the OVL.

For example, if $\alpha_1 = 1.5$, $\beta_1 = 4.0$, $\alpha_2 = 3.0$ and $\beta_2 = 1.5$ (see Figure 3) then the OVL can be computed as follows. We first equate the two distributions and use Newton's method to determine the "crossing point" which is $x = 0.56771$. Next, we use Equation 1 and the CDF of the Weibull distribution to compute the value of the OVL as follows:

$$\text{OVL} = F_1(0.56771) + [1 - F_2(0.56771)] = 0.14428 + 0.27714 = 0.42142. \tag{14}$$

OVL Between Two Beta Distributions

Here the OVL between two beta distributions is computed. The density of the beta random variable is

$$f(x) = \frac{\Gamma(p+q)}{\Gamma(p) + \Gamma(q)} x^{p-1} (1-x)^{q-1} \qquad 0 < x < 1 \quad p , q > 0 \ . \tag{15}$$

The two densities will intersect at one or two points depending on the shape and scale parameters used. These points can be found by setting the two densities equal and using Newton's method. Once the point(s) of intersection are found, Equation 1 can again be used to compute the value of the OVL. For example, if $p_1 = 2.0$, $q_1 = 2.0$, $p_2 = 1.0$ and $q_2 = 1.0$,

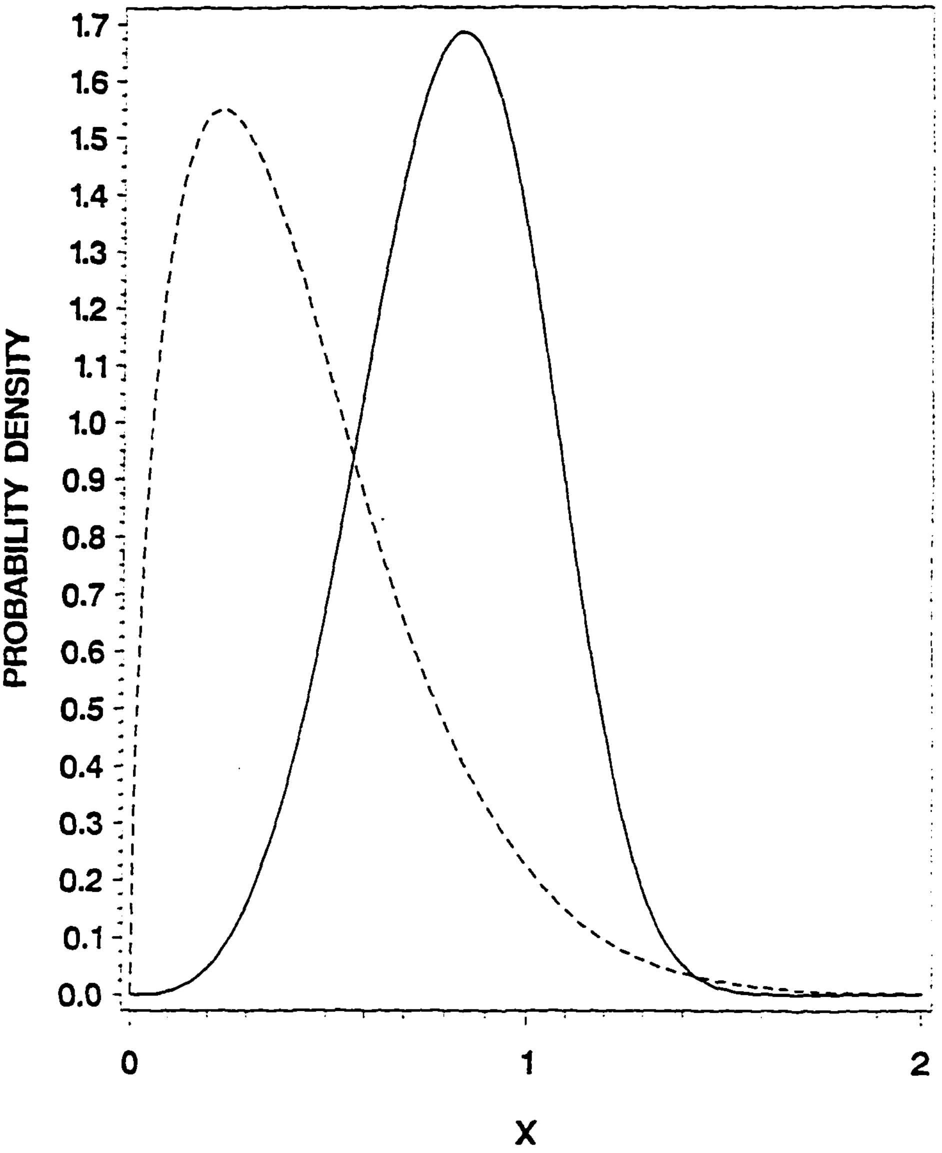

Figure 3. The overlap between two Weibull distributions. The solid line denotes a weibull distribution with $\alpha = 1.5$ and $\beta = 4.0$. The dotted line denotes a weibull distribution with $\alpha = 3.0$ and $\beta = 1.5$.

then the OVL can be computed as follows. As shown in Figure 4, the two densities intersect at two points. These points were found by setting the two densities equal and using Newton's method to determine the points of intersection, x_o = 0.21111 and x_l = 0.78890. Next we use Equation 1 to find the value of the OVL as follows:

$$\begin{aligned} \text{OVL} &= F_1(0.21111) + [F_2(0.21111)-F_1(0.78890)]+[1-F_2(0.78890)] \\ &= 0.11489 + 0.57779 + 0.11487 = 0.80755 \qquad (16) \end{aligned}$$

Previous Work Related to the OVL

Weitzman (1970) was one of the first to work with the OVL. His research included work with the discrete case of the OVL to analyze differences in income distributions of African-Americans and Whites in the United States. Gastwirth (1975) discussed several properties of the OVL. He criticized the use of the OVL as a measure of association because of its inability to detect changes in the location of the common probability mass shared by the two distributions being compared. Weitzman found the OVL to be inferior to other measures of association including the Mann-Whitney form of the Wilcoxon test for equality of population means.

Other investigators have published material using the concept of the overlap of distributions in unrelated contexts. Marx (1976) developed the overlapping coefficient as a measure of association between two normal distributions with equal variance. His development comes close to the form developed by Bradley and Piantadosi. Marx mistakenly relies on the relationship of a sample estimator of the overlap between two identical normal distributions to the central t distribution to produce a table of critical values for the sample overlapping coefficient.

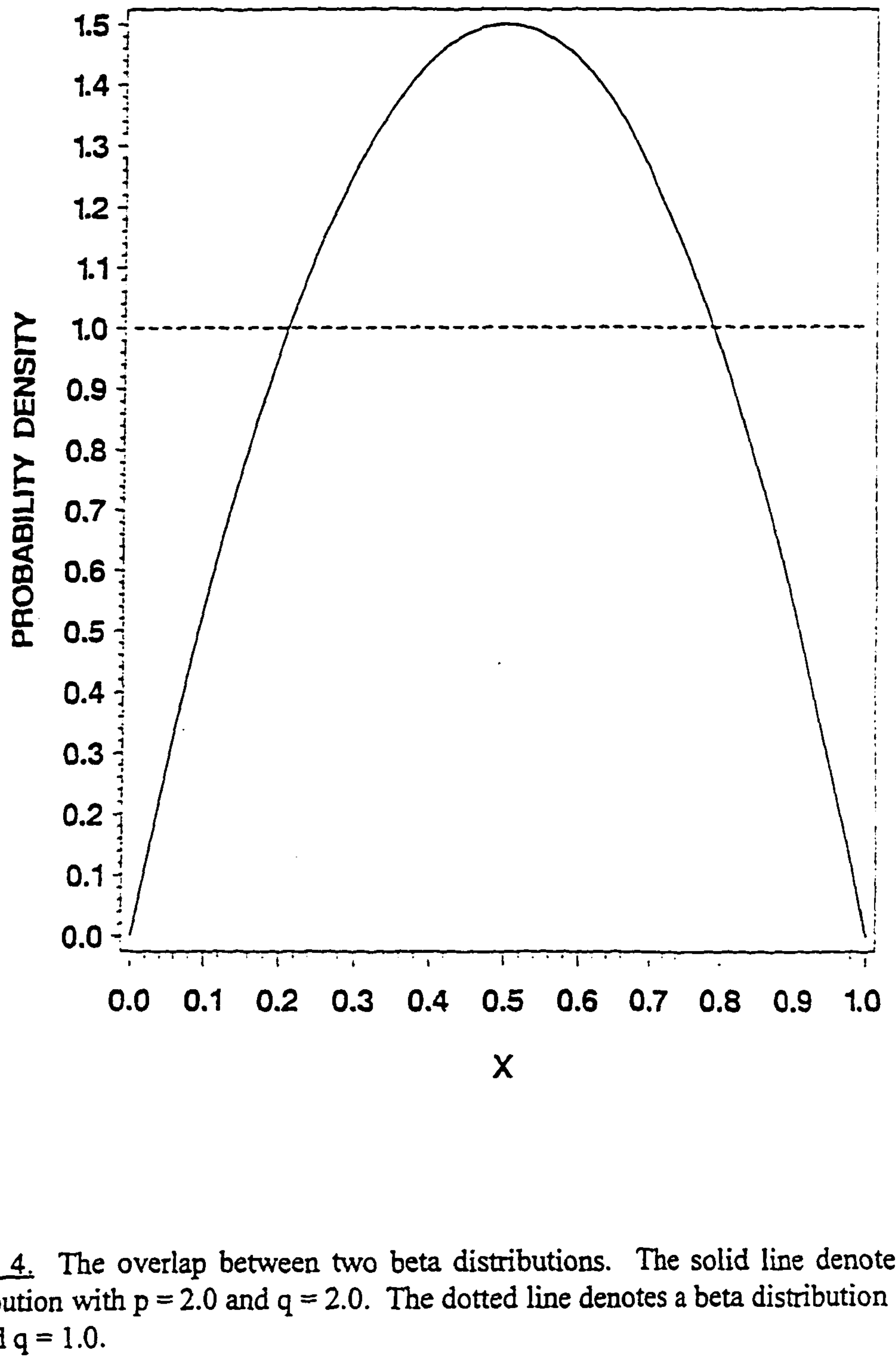

Figure 4. The overlap between two beta distributions. The solid line denotes a beta Distribution with p = 2.0 and q = 2.0. The dotted line denotes a beta distribution with p = 1.0 and q = 1.0.

He also assumes that because the sample realizations of the OVL must lie between zero and unity, the sample overlapping coefficient can be treated as the usual sample of a population proportion. Sneath (1977, 1979) used the concept of the overlap in the context of cluster analysis. He developed a method for testing the distinctness of two clusters in Euclidean space.

Bradley and Piantadosi (1982) re-introduced the OVL as a valid measure of association. They showed that the OVL was a useful method of determining the meaningfulness of an estimated difference between two probability distributions of any form. Bradley and Piantadosi derived the OVL for two normal distributions having equal and unequal variances. Mishra, Shah, and Lefante (1986) generalized the two group t-test to produce a hypothesis testing procedure and confidence intervals on the OVL of two normally distributed populations with common variance. The hypothesis testing procedure and the associated confidence limits were found to be flawed and were criticized in Inman and Bradley (1994).

Inman (1984) investigated the properties and sampling behavior of the OVL when sampling from two discrete distributions which are arranged in a 2 x C contingency table. Estimates of the sampling variance were also derived. It was found that the estimator of the OVL performed well. The estimator exhibited a downward bias (i.e., the value of the OVL is under-estimated). Also, the bias increased as the similarity of the distributions from which the samples were obtained increased.

Inman also expanded on the work of Bradley and Piantadosi. He developed a maximum likelihood estimator and an approximate variance formula for the OVL for two normal distributions with equal variances. He examined the sampling behavior of this estimator of the OVL. It was found that the sample estimators of the OVL had a downward bias trend which increases as the similarity of the distributions of interest increases. Inman and Bradley (1994) reviewed conditional tests of hypothesis and constructed direct tests of hypothesis for the true overlap for the case of the OVL for two normal distributions with equal variances. Their paper also included a method of constructing exact confidence intervals for the true overlap, along with several alternative methods of obtaining confidence intervals.

Inman also briefly looked at the case of two normal distributions with unequal variances. He assumed that the variance of the second normal distribution was larger than the variance of the first normal distribution. By equating the two probability functions he found solutions, via the quadratic formula, for the intersection of the two probability functions. Using these points he developed a maximum likelihood estimator of the OVL and an approximate variance formula. Again the bias exhibited a downward trend and also increased as the similarity of the distributions increased. Mishra and Mulekar (1992) developed confidence limits for the OVL for two normal distributions with unequal variances conditioned on the variances of the distributions which have been criticized. Clemons (1996) reparameterized the OVL for two normal probability distribution functions with unequal variances. It was found that the re-parameterization of the OVL greatly eased the computation and evaluation of the OVL. A maximum-likelihood estimator for this re-parameterization was developed. Yet the bias associated with the estimator was large for small sample sizes (i.e., $n \leq 50$). Also the bias was largest when the two distributions were similar. Yet the bias greatly decreased as the two distributions became more distinct. For the case of similar distributions, it was recommended that Inman's limiting case for equal variances be used. An approximate variance formula for the OVL was also developed. It was found that the approximate variance tended to over-estimate the variance of the OVL with the bias again being greatest for small sample sizes. Clemons also examined the performance of an asymptotic confidence interval for the reparameterized OVL.

II. NONPARAMETRIC ESTIMATION OF THE OVL

It is sometimes the case that data collected suggest no reasonable parametric form for $f_1(x)$ or $f_2(x)$ or both. There are several approaches for the estimation of the OVL for such a circumstance . First, one could try transforming the data and estimate the OVL using the invariance property of the OVL. Second, one could use a "quasi-parametric" approach, using a flexible family of distribution functions, such as Pearson, Burr, or Johnson families of distributions (Johnson and Kotz, 1970, pp. 9-33), to characterize the two distributions. Using the characterization of the two distributions, one can estimate the value of the OVL. The last approach would be to estimate the two distributions nonparametrically, using one of several available nonparametric density estimation procedures (Wegman, 1972, 1982).

The focus of a nonparametric density estimator is to obtain a good estimate of the density function with minimum assumptions. Nonparametric techniques are used because they eliminate the need to specify a form of the model. The disadvantage of the nonparametric techniques is that these techniques result in a loss of efficiency. Yet, the loss of efficiency is balanced by the reduction of the risk of misinterpreting the data by incorrectly specifying the parametric form of the function.

Inman (1984) examined the properties of an estimator of the OVL when sampling from two distributions estimated nonparametrically by quadratic splines. By fitting quadratic spline functions to the empirical distribution through weighted least squares and taking the derivatives of these spline functions as the estimated densities and using the density estimates to determine points of intersection, he obtained a nonparametric estimate of the OVL between $f_1(x)$ and $f_2(x)$. Inman then used a bootstrap variance estimate to compute the variance of the estimated OVL. Through a Monte Carlo study, he found that the nonparametric estimator of OVL performed well as an estimator of OVL. It was found that the estimator was a biased estimator of the OVL. The estimator generally under-estimated the true overlap. The bias of the estimator was found to be related to the value of the OVL and the sizes of the two samples. Inman suggested that because of the success of the spline-density based technique of estimating the OVL, a less sophisticated nonparametric method might prove adequate in settings where the distributional assumptions seem unwarranted. One alternative is the naive kernel estimate (Rosenblatt, 1956; Waterman & Whiteman, 1978), which will be examined in this study.

To learn something of the properties of the kernel estimator of the OVL, it is compared to the true overlap for several known distributions, the normal, gamma, beta, and Weibull distributions, via a Monte Carlo study. Also the nonparametric estimator is compared to the maximum likelihood estimator of the OVL (Inman, 1984; Clemons, 1996) for the normal distribution case.

Kernel Density Estimation

Until the 1950s the histogram was the only nonparametric density estimation technique (Scott, 1992). Fix and Hodge (1951) introduced an algorithm for computing a nonparametric density estimator by exploring the statistical discrimination when the parametric density is unknown. Rosenblatt (1956) developed a general form of estimating a density function nonparametrically using what is called the naive or Rosenblatt kernel density estimator.

The basic idea of the naive kernel estimator is adopted from the idea behind the basic histogram. The kernel estimator uses the empirical density function which is a histogram-type estimate of the underlying density function that is fairly easy to compute and understand. The empirical density function is a simple modification of the histogram and has convergence properties that are equivalent to those of the larger class of estimators (Waterman & Whiteman, 1978). If we have an unknown density of a continuous random variable from sample data, $x_1, ..., x_n$, which are independent and identically distributed with distribution function F(x), the empirical distribution function is defined as

$$F_n(x) = \frac{number\ (X_i : X_i < x)}{n} \ . \tag{17}$$

By an application of the binomial distribution (Hogg & Craig, 1970) it is has been shown that

$$\lim_{n \to \infty} F_n(x) = F(x) \tag{18}$$

with probability 1. The kernel density is a numeric approximation of the derivative of the emperical cumulative distribution function. Using the fact that dF(x)/dx = f(x), the approximate derivative of $F_n(x)$, the Rosenblatt estimator (naive kernel estimator) is given by

$$\hat{f}_n(x) = \frac{F_n(x + h) - F_n(x-h)}{2h} , \tag{19}$$

where $h > 0$ is a real valued number constant which is a function of the sample size and approaches zero as $n \to \infty$. This h is also known as the shaping parameter, window width, or bandwidth of the kernel estimator.

The Rosenblatt kernel estimator is constructed by placing a rectangle of width $2h$ and height $(2nh)^{-1}$ on each observation and then summing to obtain the estimate. Kernel density estimators inherit all properties of the kernel. Thus, since the naive kernel is discontinuous the resulting estimate is also discontinuous. The discontinuity follows from the naive kernel estimator having "bumps" at the points $x_i \pm h$ and zero derivatives elsewhere. This results in a rough, jagged estimator.

In general, the basic kernel estimator is of the form

$$\hat{f}_n = \int_{-\infty}^{\infty} K_n(x,y)dF_n(y) = \sum \frac{1}{n}\sum_{i=1}^{n} K_n(x,x_i)\,, \tag{20}$$

where K_n is the kernel. To adapt the naive kernel density function to the above definition we take K_n to be as follows:

$$K_n(x\,,y) = \frac{1}{2h} \quad for \ |x - y| \le h \quad and \quad zero \quad elsewhere\,. \tag{21}$$

The estimator $\hat{f}_n$ is dependent on the data as well as on the kernel specified and the bandwidth.

Bias and Variance of the Kernel Estimator

Since the bias of the kernel estimator depends on the value of the bandwidth, and bandwidth is a function of the sample size, we can also say that the bias depends indirectly on the sample size. The bias of the kernel estimator is expressed as follows (Silverman, 1986, p. 39):

$$bias_h(x) = E\hat{f}x - f(x) = \int h^{-1}K[(x-y)/h]f(y)dy - f(x)\,. \tag{22}$$

If we make a change of variable, $y = x - ht$ and use the following assumptions about the kernel, K,

$$\int K(t)dt=1\ , \quad \int tK(t)dt=0, \quad and \quad \int t^2K(t)dt=k_2\neq 0\ ; \tag{23}$$

then the bias can be expressed as follows:

$$bias_h(x) = \int K(t) f(x-ht)\,dt - f(x) = \int K(t)\,[f(x-ht) - f(x)]dt. \tag{24}$$

By using a Taylor's expansion we can express the term f(*x*-*ht*) as

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) + \dots . \tag{25}$$

Using the assumptions on K shown in equation 23, the bias can be written as

$$bias_h(x) = -hf'(x)\int t\,K(t)\,dt + \frac{1}{2}h^2f''(x)\int t^2\,K(t)dt + \cdots = \frac{1}{2}h^2f''(x)k + O(h). \tag{26}$$

The variance of the estimator can be found as follows (Silverman, 1986, p. 39-40):

$$var\,\hat{f}(x) = n^{-1}\int h^{-2}K[(x-y)\,h^{-1}]^2 f(y)\,dy - n^{-1}[f(x) + bias_h(x)]^2 . \tag{27}$$

Using Equation 26 and substituting $y = x\text{-}ht$ into the above equation we obtain the following:

$$var\,\hat{f}(x) \approx n^{-1}h^{-1}\int f(x-ht)\,K(t)^2dt - n^{-1}[f(x) + O(h^2)]^2. \tag{28}$$

Again, if we expand f(*x*-*ht*) into a Taylor's series then

$$var\,\hat{f}(x) \approx n^{-1}h^{-1}\int[f(x) - htf'(x) + \cdots]K(t)^2dt + O(n^{-1}) = n^{-1}h^{-1}f(x)\int K(t)^2dt + O(n^{-1}). \tag{29}$$

Thus, simplification of the variance of the estimator is given as follows:

$$var\,\hat{f}(x) \approx n^{-1}h^{-1}f(x)\int K(t)^2dt. \tag{30}$$

Choice of the Bandwidth of the Kernel Estimator

As shown above, the choice of the bandwidth, *h*, is what drives the kernel estimator. Since the kernel estimator has been shown to be a biased estimator, the criterion for optimization is the mean integrated square error. The mean integrated square error is one of the most widely used methods of placing a measure on the global accuracy of density estimators. Thus, the ideal value of h is that value which minimizes the approximate mean integrated square error ,

$$\int bias_h(x)^2 dx + \int var\,\hat{f}(x)\,dx = \frac{1}{4}h^4 k^2{}_2 \int f''(x)^2 dx + n^{-1}h^{-1}\int K(t)^2 dt\,. \tag{31}$$

It can be shown by calculus (Parzen, 1962) to be equal to

$$h_{opt} = k^{-2/5}\left[\int K(t)^2 dt\right]^{1/5}\left[f''(x)^2 dx\right]^{-1/5} n^{-1/5}\,. \tag{32}$$

We see that the optimum bandwidth depends on the unknown density function. Since we are assuming that the density is unknown, it is unlikely that we will know enough to choose the optimum h. The problem now becomes how to choose an efficient smoothing parameter.

Silverman (1986) states that a natural method for choosing the smoothing parameter is to plot out several curves and choose the estimator that is most in accordance with one's prior ideas about the density. This is called the subjective choice of a smoothing parameter. Secondly, one could choose h by using a standard family of distributions to assign a value to the term $\int f''(x)^2 dx$ in equation 32. Scott (1992) used this second approach and the normal distribution as the parametric family to obtain the normal reference rule bandwidth

$$h = \left(\frac{4}{3}\right)^{1/5} \sigma\, n^{-1/5} = 1.06\,\sigma\, n^{-1/5}\,. \tag{33}$$

Thus a simple way of choosing the smoothing parameter would be to estimate σ from the data and substitute into Equation 33 . This method works well if the population is truly normal, yet may tend to over-smooth if the population is multimodal since $(\int f''(x)^2)^{-1/5}$ is large relative to the standard deviation. Silverman suggests that a better result may be obtained by using the interquartile range, R. Yet, if the underlying distribution is bimodal, using the interquartile range tends to over-smooth even further. It is then suggested to use an adaptive estimate of spread, A = min (standard deviation , interquartile range / 1.34) , instead of σ in Equation 33. Using this normal (adaptive) reference rule is attractive in that it is a fully automatic method of choosing a smoothing parameter. It also allows researchers reporting results a reference to a standardized method of estimating the shaping parameter.

Estimation of the OVL With Kernel Estimates

Given the procedure for estimating the unknown density using a kernel estimator developed above, obtaining the estimate of OVL, $O\tilde{V}L$, will be approached as follows. FORTRAN subroutines (Appendix A) were used to compute the kernel estimator of the overlapping coefficient. The compilation and execution of the programs were performed on the Cray C-90 supercomputer.

From two independent samples from unknown distributions, $x_{11}, \ldots, x_{1n_1}$ and $x_{21}, \ldots, x_{2n_2}$, we compute the density estimates using the FORTRAN subroutine OVCOEF, which uses Equation 19 in conjunction with the Alternative Reference Rule for obtaining a value of the bandwidth for the formulation of the density estimates. Once the density estimates are computed, the value of the overlapping coeffiecient is computed by finding the jump points (i.e., the points where the density of the kernel density estimator changes) using the FORTRAN subroutine JUMPS. The jump points for each sample are then combined into one set of points and then sorted using the IMSL (1991) FORTRAN subroutine VSRTD. The intervals between consecutive points are computed using the FORTRAN subroutine INTERV (de Boor, 1978). Finally, the $O\tilde{V}L$ is computed by summing the area under the smaller curve over each subinterval.

Consider for example the two kernel estimated densities in Figures 5 and 6, which are obtained from two samples of size 500 generated from two normal distributions. The

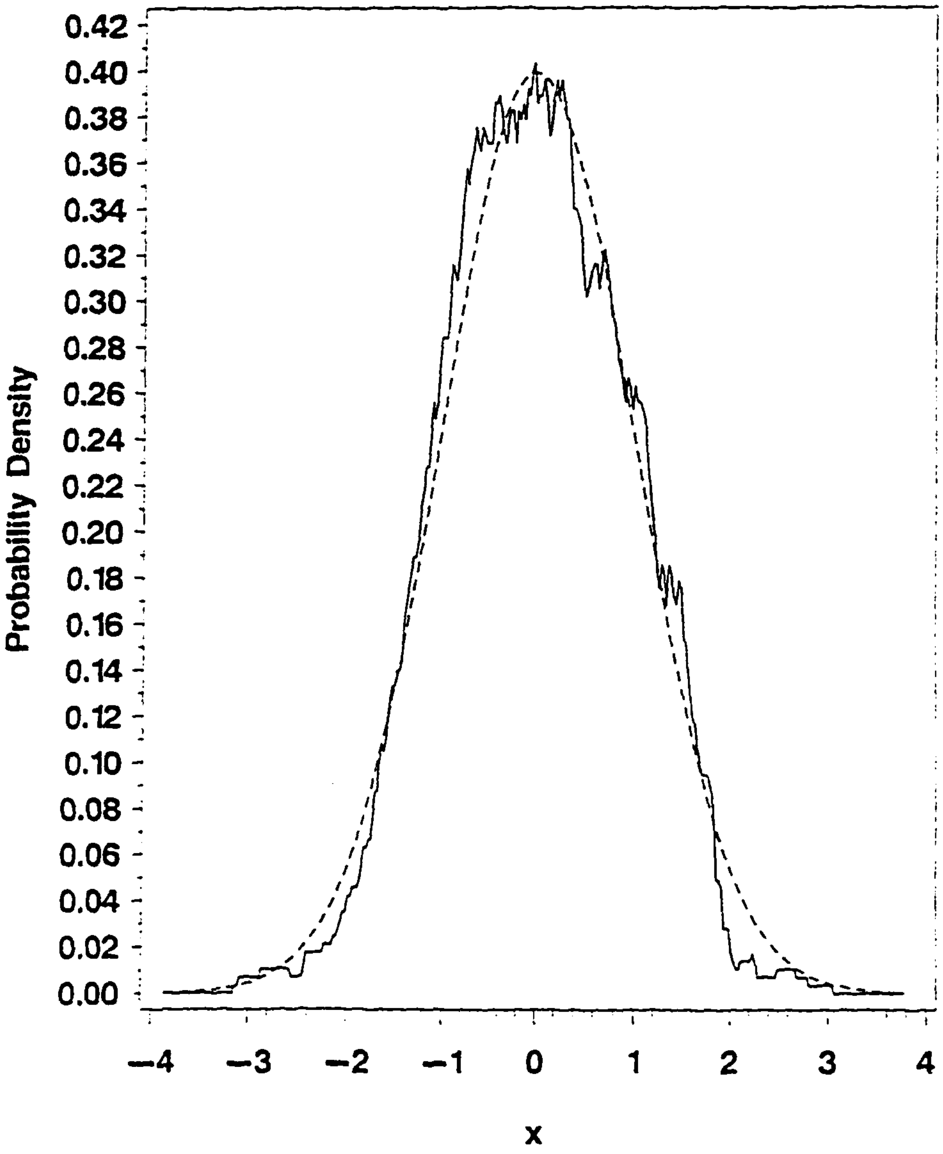

Figure 5. The kernel density estimator for a standard normal distribution. The solid line denotes a kernel density generated from pseudo normal random deviates using the normal reference rule. The dotted line denotes a standard normal distribution.

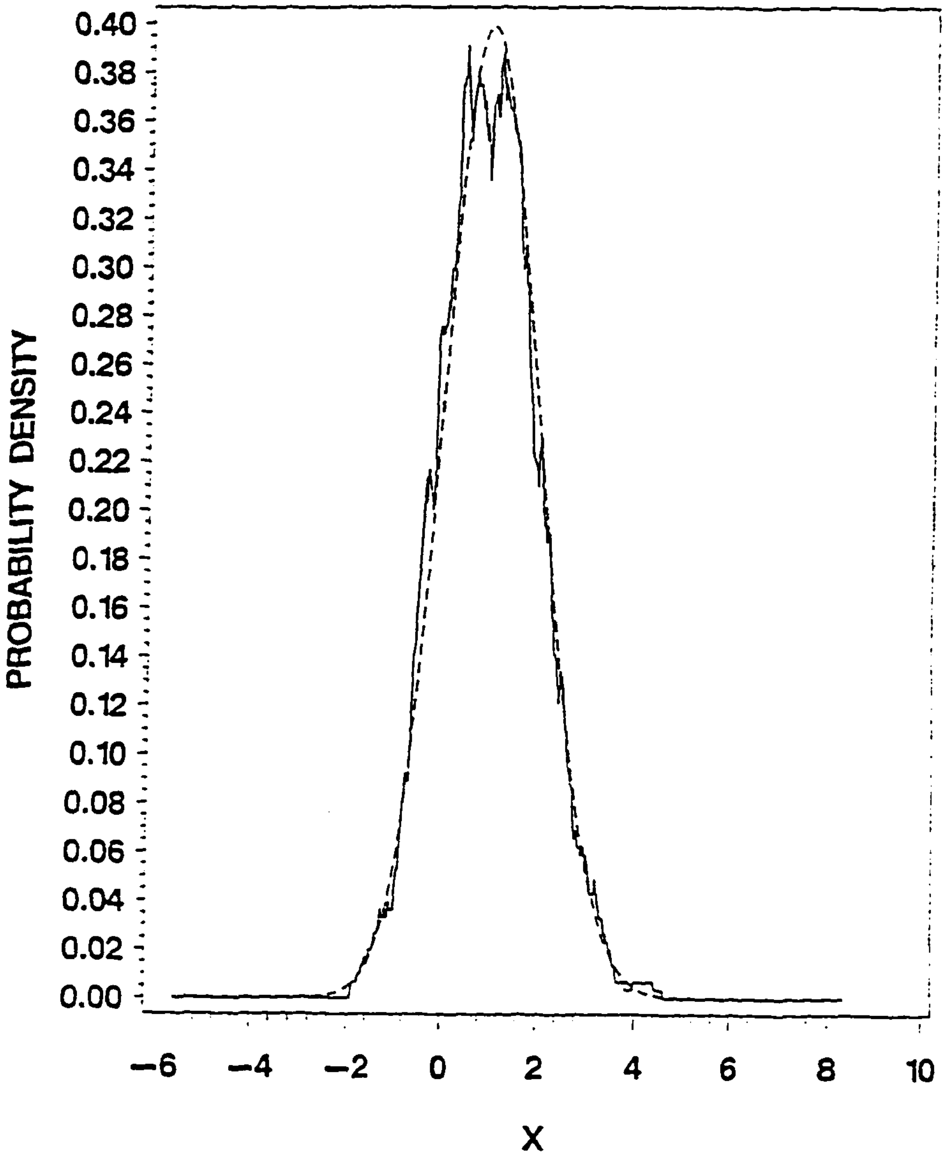

Figure 6. The kernel density estimator for a normal distribution. The solid line denotes a kernel density generated from pseudo normal random deviates using the normal reference rule. The dotted line denotes a normal distribution with mean = 1 and variance = 1. first

first sample is generated from the standard normal distribution; the density estimate derived from this sample is indicated by a dotted line. The second is from a normal distribution with mean of 1 and variance of 1 . Using the subroutines described previously we find that the $O\tilde{V}L$ is 0.63601374 (see Figure 7). The actual OVL (see Figure 8) between the two normal distributions is 0.617075.

Nonparametric Estimator of the Variance

An alternative to the approximation of the variance of the overlapping coefficient may be achieved by using a nonparametric estimation approach, the bootstrap, (Efron, 1979, 1981, 1982; Efron & Gong, 1983; Efron & Tibshirani, 1986, 1993). The bootstrap is one of the simplest nonparametric variance estimation techniques available. It was introduced by Efron (1979) as a computer-based method for estimating the standard error of a random variable. The bootstrap algorithm works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of the random variable by the empirical standard deviation.

The basic idea for a bootstrap estimator of variance is as follows. Suppose we are given two independent samples of $x_{11}, \dots, x_{1n_1}$ and $x_{21}, \dots, x_{2n_2}$. By treating the samples as two finite populations of size n_1 and n_2 respectively, we can draw with replacement two new bootstrap samples each of the size of the original samples. Thus we have what is known as pseudo-data, $x_{11}^{*}, \dots, x^{*1n_1}$ and $x_{21}^{*}, \dots, x_{2n_2}^{*}$. Using this pseudo data, we then calculate the value of the kernel estimated OVL, $O\tilde{V}L$. This resampling procedure is repeated some large number, say B, of times with a new bootstrap sample being generated each time.

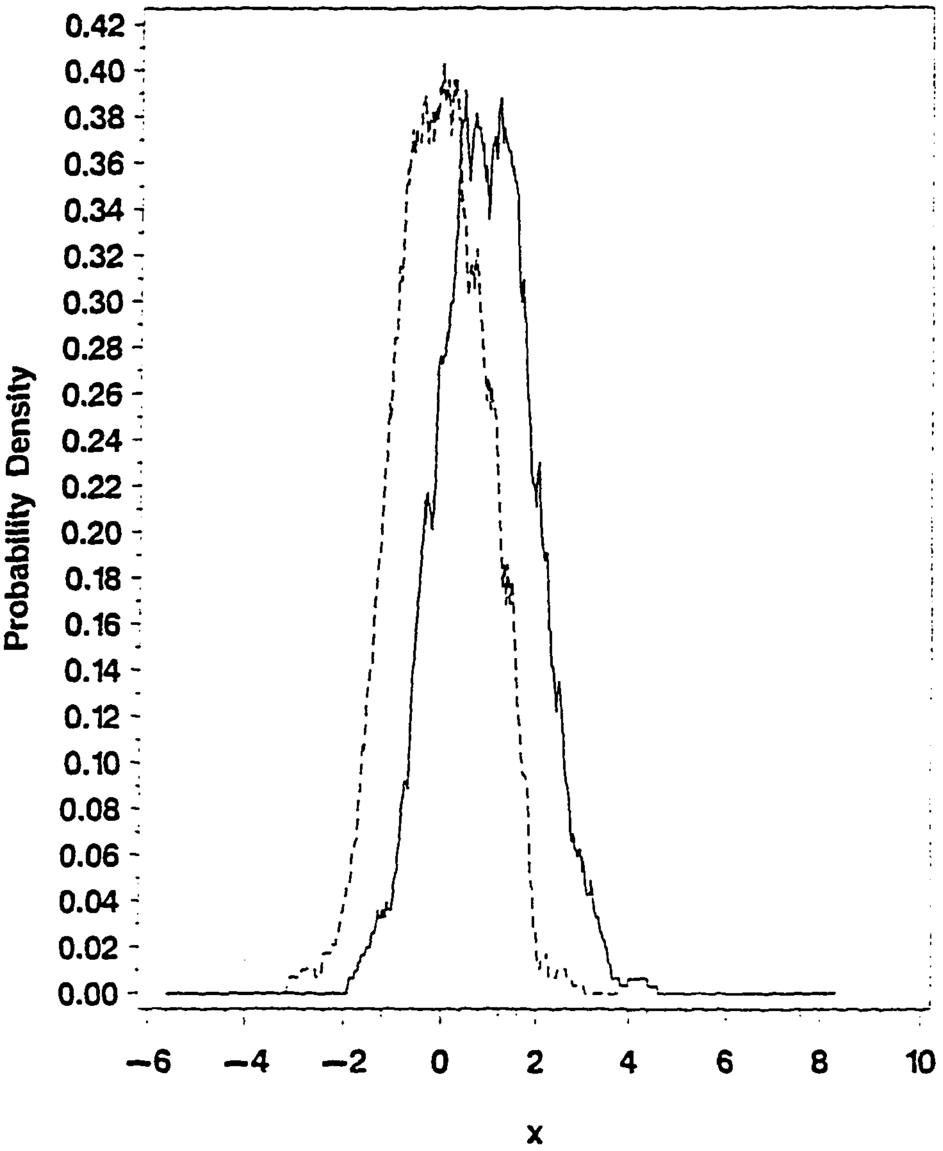

Figure 7. The overlap using the kernel densities for two normal distributions. The solid line denotes a kernel density generated from pseudo normal random deviates with mean = 1 and variance = 1 using the normal reference rule. The dotted line denotes a kernel density generated from pseudo standard normal random deviates.

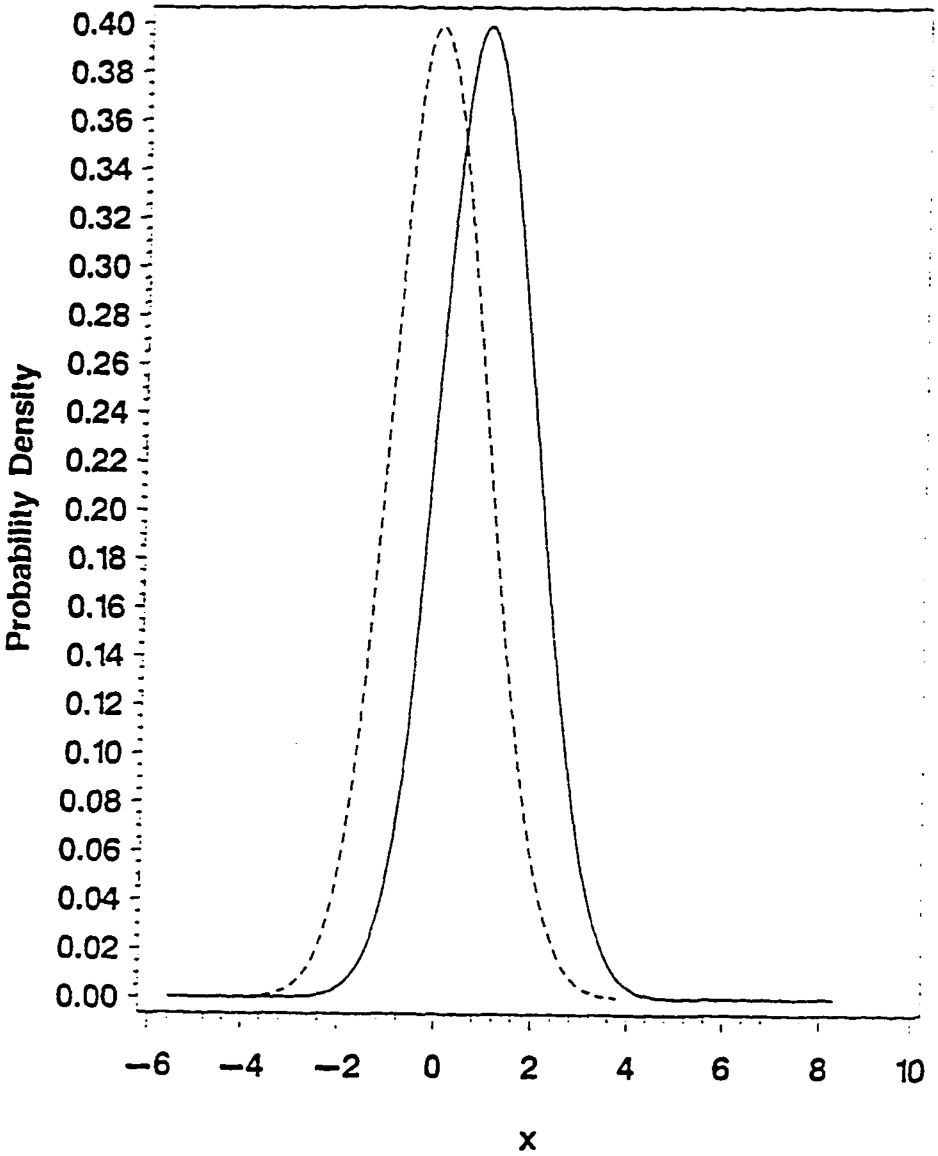

Figure 8 . The "true" overlap for two normal distributions. The solid line denotes a normal distribution with mean = 1 and variance = 1. The dotted line denotes a standard normal distribution.

If we let $O\tilde{V}L_i^*$ denote the value of $O\tilde{V}L$ computed on the ith iteration, then the bootstrap estimator of the variance of $O\tilde{V}L$ is given by:

$$Var\tilde{O}VL_B(O\tilde{V}L) = \frac{\sum_{i=1}^{B}(O\tilde{V}L_i^* - O\tilde{V}L^*)^2}{B-1} \tag{34}$$

where

$$O\tilde{V}L^* = \frac{\sum_{i=1}^{B} O\tilde{V}L_i^*}{B}. \tag{35}$$

The only difficulty in computing this new estimate is the value of B. According to Efron and Tibshirani (1993, p.52) a small number of bootstraps, say B = 25, can be considered informative. A value of B = 50 is often enough to give a good estimator. Very seldom are more than B = 200 replications needed for estimating, yet a much larger value of B is usually required for bootstrap confidence intervals.

Confidence intervals can be constructed using the percentile method for bootstrap variance estimates (Efron, 1982). Let $F_B^*(\cdot)$ be the empirical distribution function constructed from the bootstrap estimates of the OVL (i.e., $O\tilde{V}L_i^*$ (i = 1, ..., B)) and $F_B^{*\,-1}(\cdot)$ denote its inverse. A (1-α)100% confidence interval for OVL is using the percentile bootstrap variance method is as follows:

$$\left(F_B^{*\,-1}(\alpha/2), F_B^{*\,-1}(1-\alpha/2)\right). \tag{36}$$

Monte Carlo Investigation

To determine the properties of the kernel estimator of the OVL, $O\tilde{V}L$ has been calculated on a set of Monte Carlo samples from two normal, two gamma, two Weibull, and two beta distributions, using a selected number of design points for each distribution. This study will assess the $O\tilde{V}L$ as an estimator when sampling from two distributions are identical (i.e, OVL = 1); when the two distributions are similar (i.e., 1 > OVL > .500); and when the two distributions are quite distinct (i.e., OVL < 0.500). Since the OVL is a tool used to measure the common area between two distributions deemed by hypothesis tests to differ, this investigation is biased towards larger values of the OVL.

For the two normal distributions, the twelve design points chosen consist of combinations of the following: $\mu = 0, \sigma^2 = 1$; $\mu = 2, \sigma^2 = 4$; $\mu = 3, \sigma^2 = 5$; $\mu = 1, \sigma^2 = 1$; $\mu = 0, \sigma^2 = 3$; $\mu = 5, \sigma^2 = 10$; and $\mu = 3, \sigma^2 = 5$. The gamma distribution (see Figures 9-11) used in the simulation study will be defined as follows: $f_x(x) = \dfrac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}$ for x, $\alpha > 0$. The twelve design points evaluated for the gamma distribution case will be combinations of the following: $\alpha = 1.5$; $\alpha = 2.0$; $\alpha = 2.5$; $\alpha = 3.5$; and $\alpha = 4.0$. The Weibull distribution (see Figures 12-14) used in the study is defined as follows: $f_x(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}$ for x, α, $\beta >$ 0. The twelve design points for the Weibull distribution will consist of combinations of the following parameters: $\alpha = 1.5, \beta = 4.0$; $\alpha = 2.0, \beta = 2.0$; $\alpha = 1.5, \beta = 1.5$; $\alpha = 2.0, \beta = 3.0$; $\alpha = 1.5, \beta = 1.5$; $\alpha = 1.0, \beta = 2.0$; $\alpha = 1.0, \beta = 3.0$; and $\alpha = 1.0, \beta = 3.5$. Four design points for the beta distribution will consist of combinations of the following parameters: p = 2, q = 2; p = 1, q = 1; p = 3, q = 3; and p = 5, q = 3, where the beta distribution is defined as $f_x(x) = \dfrac{\Gamma(p+q)}{\Gamma(p)+\Gamma(q)}x^{p-1}(1-x)^{q-1^{\beta-1}}e^{-\alpha x^{\beta}}$, $0 < x < 1$ and p, q > 0. A complete list of design points and values of the OVL is shown in Appendix B. The $O\tilde{V}L$ was also investigated for two mixtures of distributional settings: Standard normal and standard Cauchy distributions (see Figures 15 and 16) and gamma distribution with $\alpha = 3$ and a chi square distribution with 4 degrees of freedom. The sample sizes used to investigate the sampling

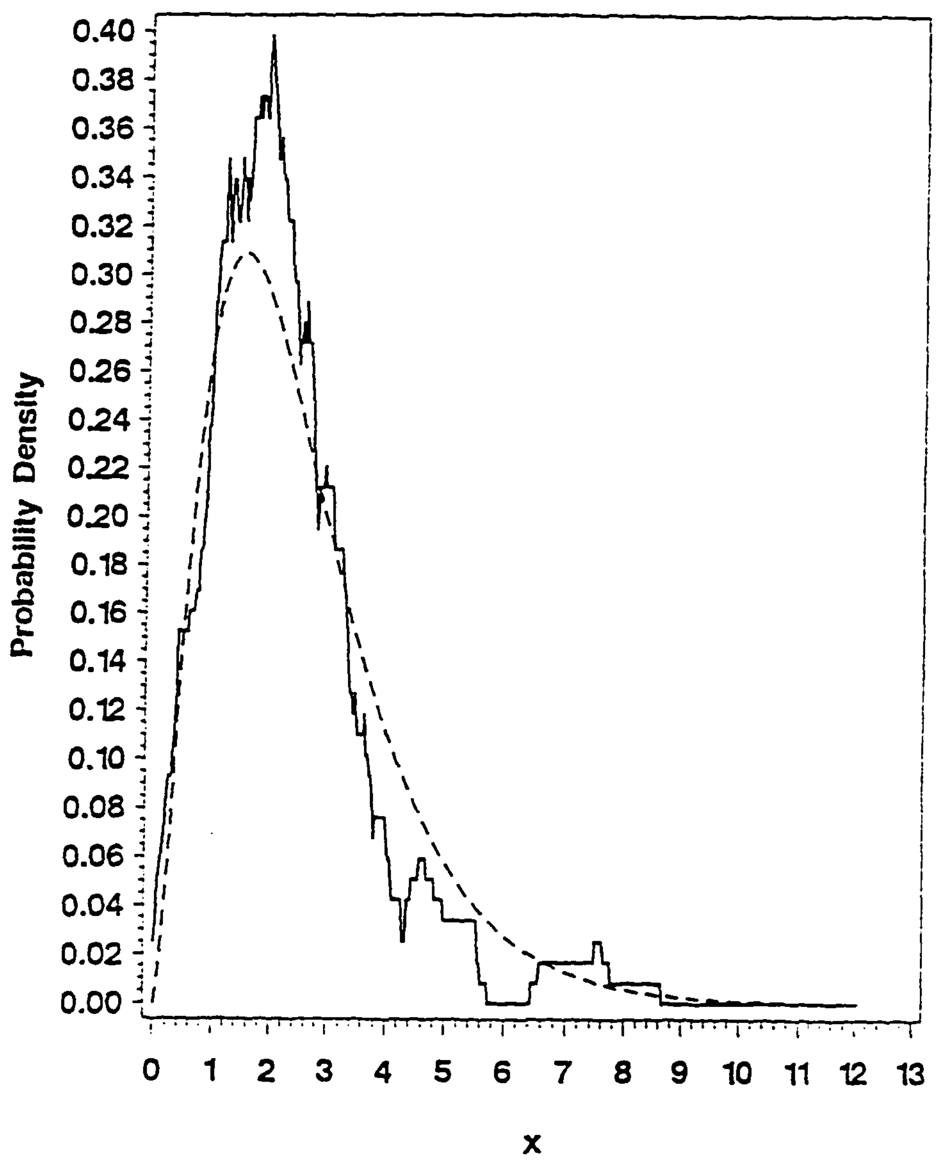

Figure 9. The kernel density estimator for a gamma distribution. The solid line denotes a kernel density generated from pseudo gamma random deviates using the normal reference rule. The dotted line denotes a gamma distribution with $\alpha = 2.5$.

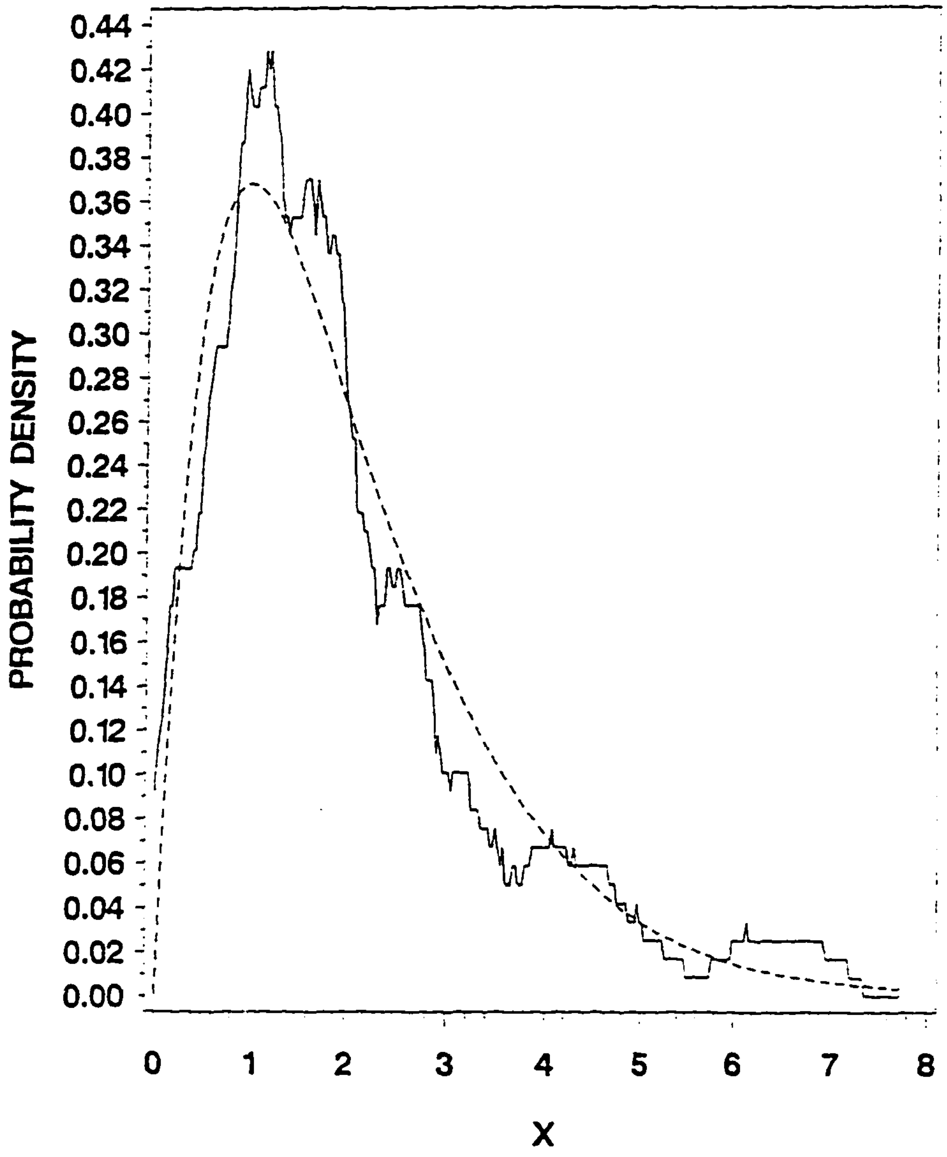

Figure 10. The kernel density estimator for a gamma distribution. The solid line denotes a kernel density generated from pseudo gamma random deviates using the normal reference rule. The dotted line denotes a gamma distribution with $\alpha = 2.0$.

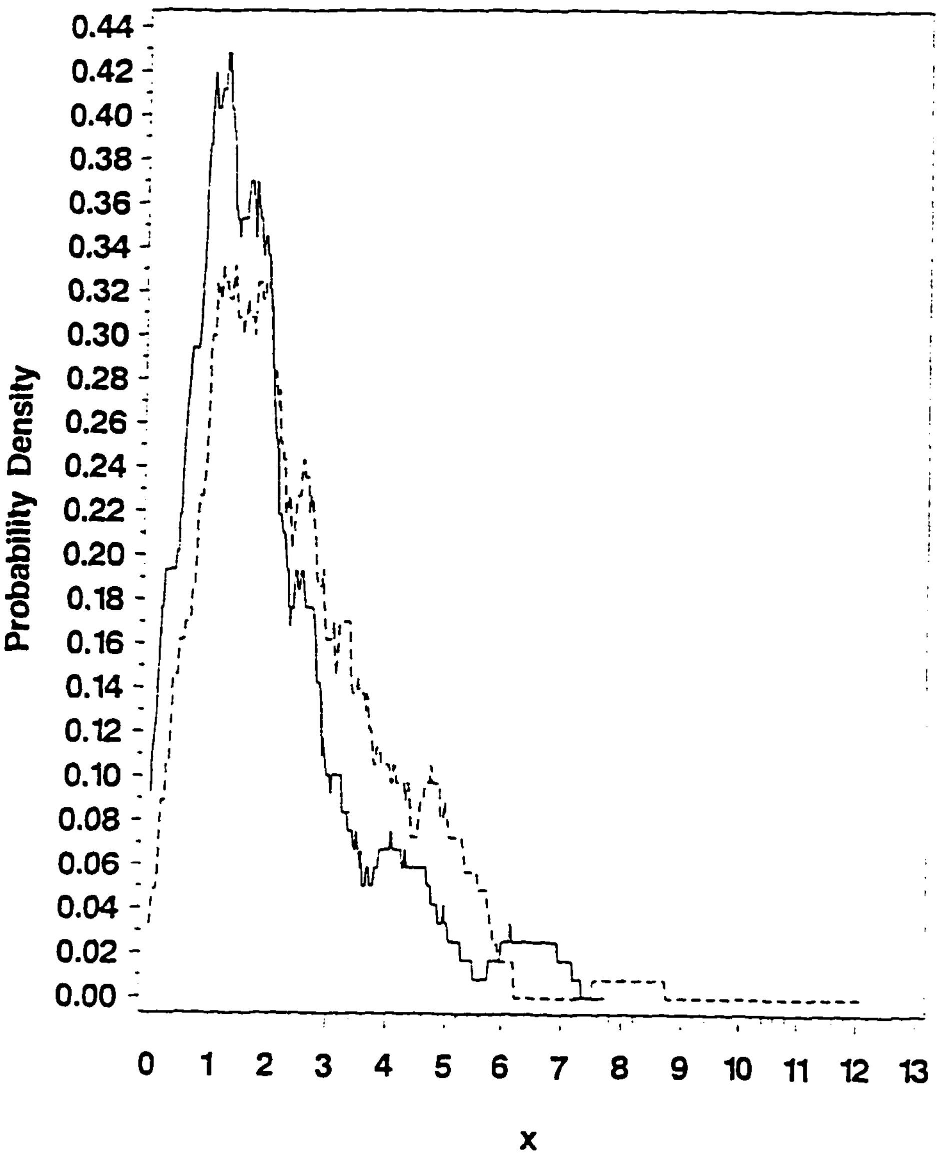

Figure 11. The overlap using the kernel densities for two gamma distributions. The solid line denotes a kernel density generated from pseudo gamma random deviates with $\alpha = 2.0$ using the normal reference rule. The dotted line denotes a kernel density generated from pseudo gamma random deviates with $\alpha = 2.5$.

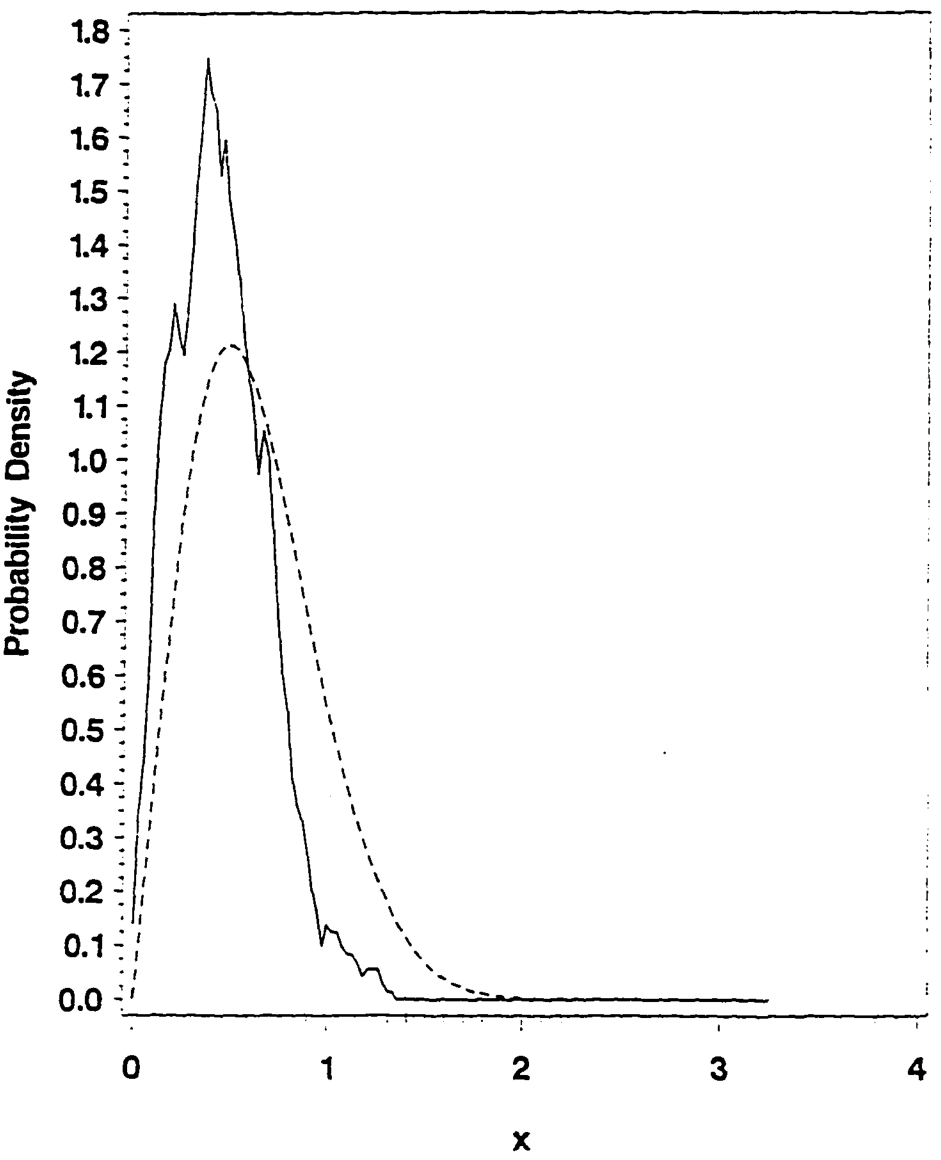

Figure 12. The kernel density estimator for a Weibull distribution. The solid line denotes a kernel density generated from pseudo Weibull random deviates using the normal reference rule. The dotted line denotes a Weibull distribution with $\alpha = 2.0$ and $\beta = 2.0$.

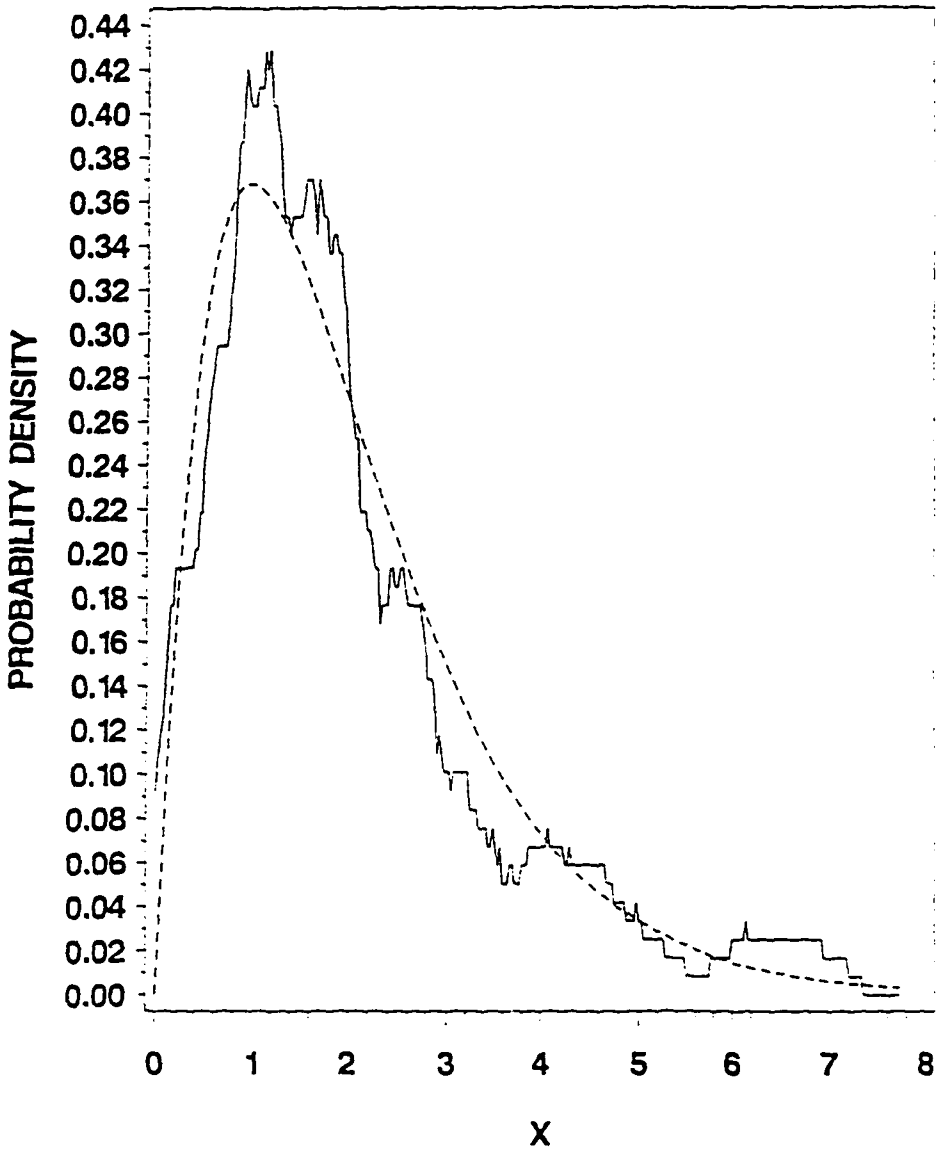

Figure 13. The kernel density estimator for a Weibull distribution. The solid line denotes a kernel density generated from pseudo Weibull random deviates using the normal reference rule. The dotted line denotes a Weibull distribution with $\alpha = 1.5$ and $\beta = 1.5$.

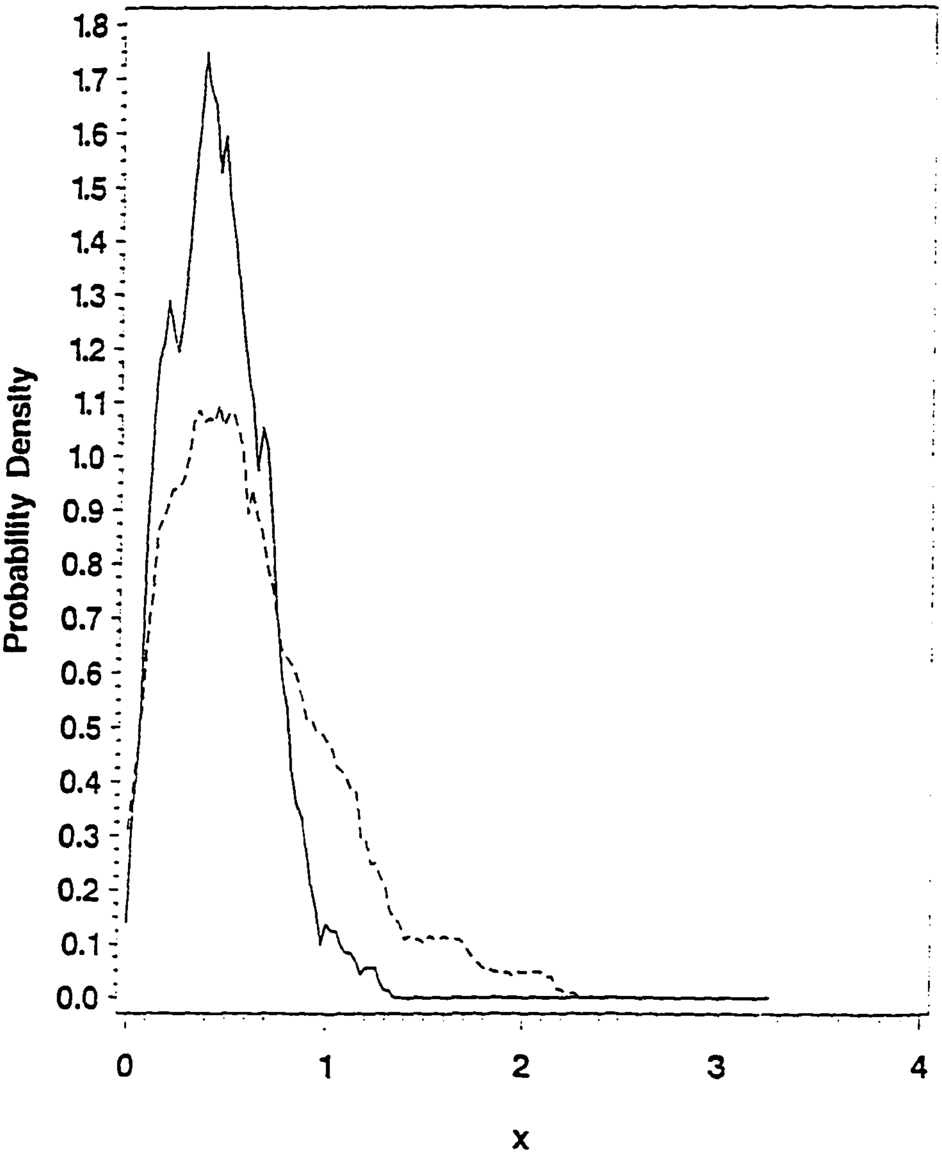

Figure 14. The overlap using the kernel densities for two Weibull distributions. The solid line denotes a kernel estimator generated from pseudo Weibull random deviates with $\alpha = 2.0$ using the normal reference rule. The dotted line denotes a kernel estimator generated from pseudo Weibull random deviates with $\alpha = 2.5$.

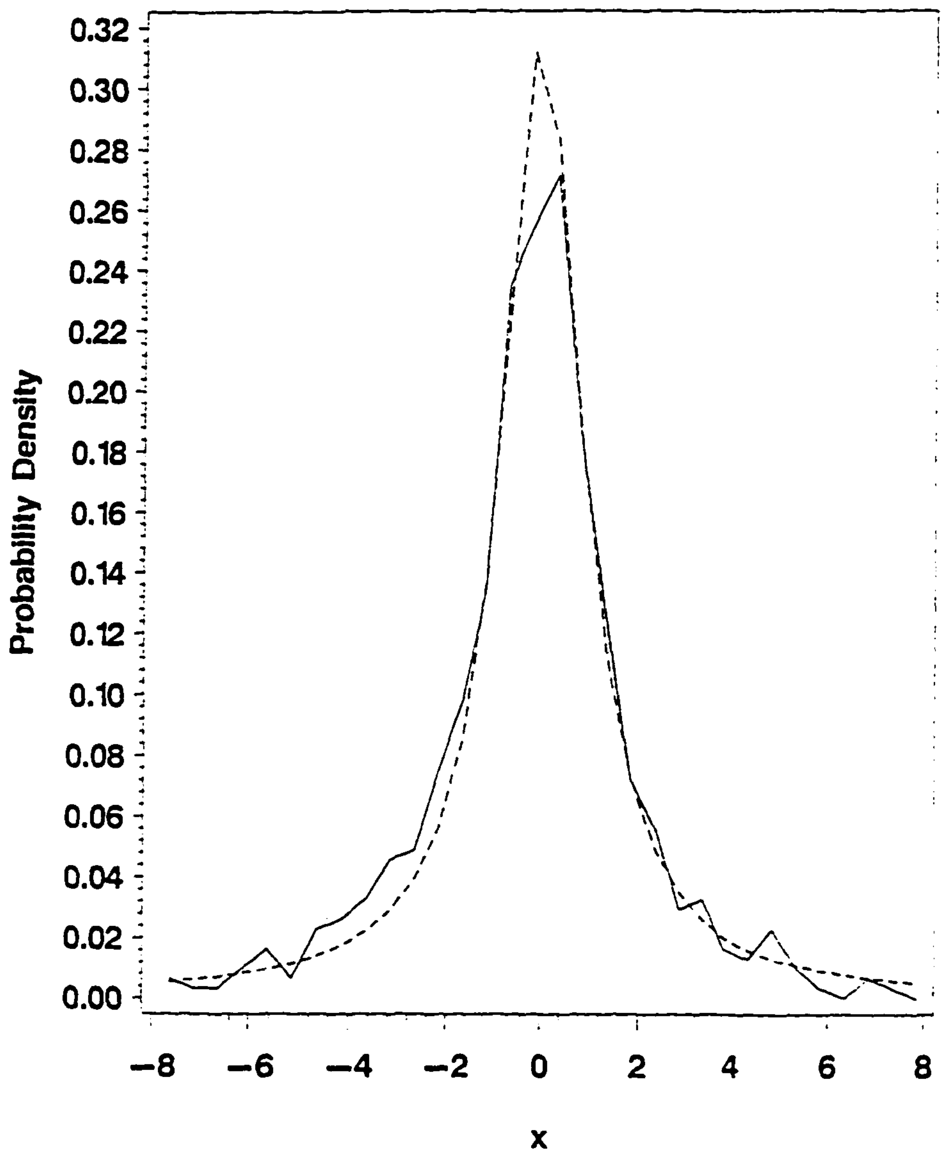

Figure 15. The kernel density estimator for a standard Cauchy distribution. The solid line denotes a kernel density generated from pseudo standard Cauchy random deviates using the normal reference rule. The dotted line denotes a standard Cauchy distribution.

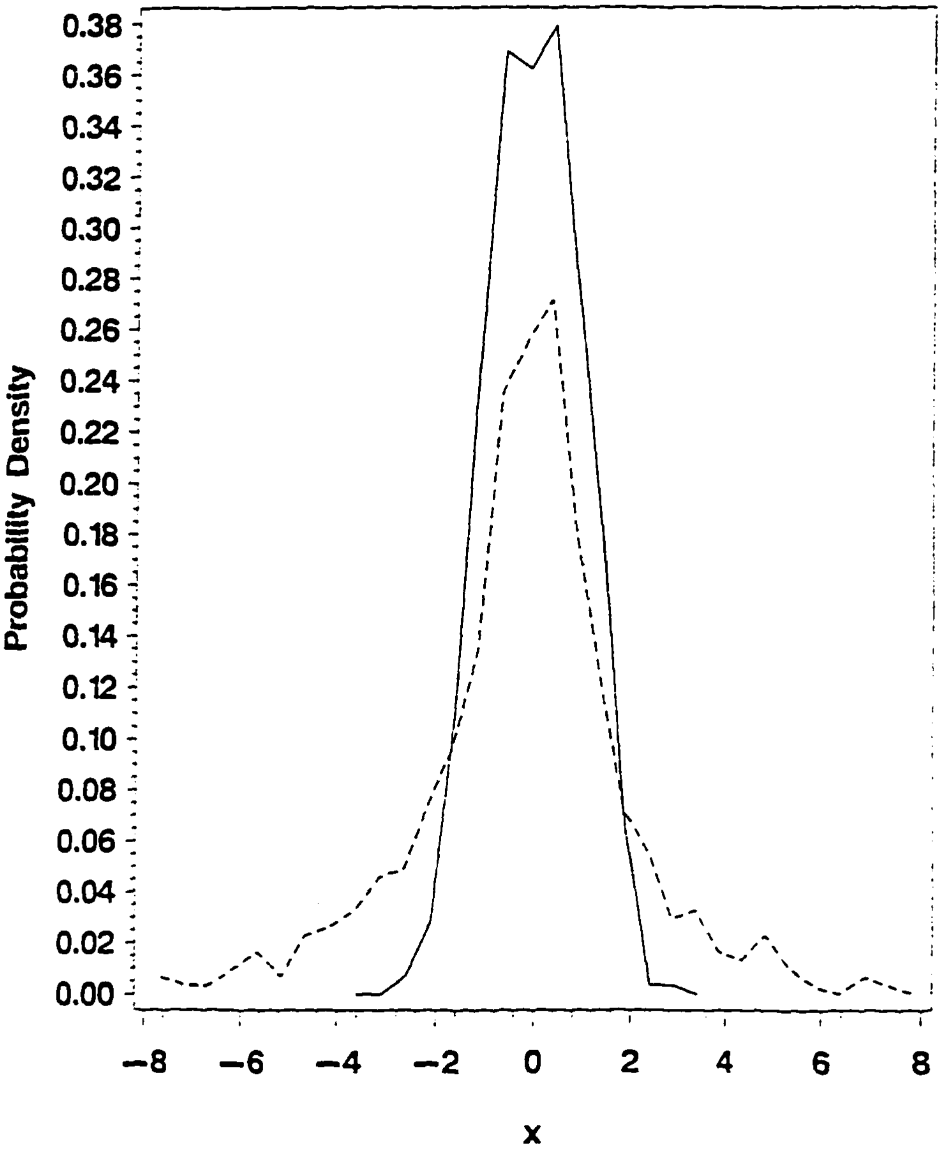

Figure 16. The overlap using the kernel densities for a standard Cauchy distribution and a standard normal distribution. The solid line denotes a kernel estimator generated from pseudo standard normal random deviates using the normal reference rule. The dotted line denotes a kernel estimator generated from pseudo standard Cauchy random deviates.

behavior of the Kernel estimator of OVL, $O\tilde{V}L$ will be $n_1 = n_2 = 100$ and 500. The random deviates were generated from the following IMSL (1991) routines:

DRNBET--generates double precision pseudo random numbers from a beta distribution

DRNCHI--generates double precision pseudo random numbers from a chi-square distribution

DRNGAM--generates double precision pseudo random numbers from a standard gamma distribution

DRNNOA--generates double precision pseudo random numbers from a standard normal distribution

DRNCHY--generates double precision pseudo random numbers from a Cauchy distribution

DRNWIB--generates double precision pseudo random numbers from a Weibull distribution.

On each of the 1000 Monte Carlo trials at each design-point-sample-size combination, $O\tilde{V}L$ is computed as described previously. To investigate the bootstrap estimator of the variance of $O\tilde{V}L$, the bootstrap estimate of variance was calculated using Equation 34 for each design-point-sample-size combinations using B = 200. The IMSL random number generator for a uniform (0,1) distribution, DRNUNF, was used to generate a random sample with replacement from the generated distributions. The bootstrap estimator was then calculated using the subroutine MEANSTA (Miller, 1982).

The Monte Carlo mean and variance were computed from the observed first and second sample moments from the simulated samples. The "true" value of the OVL was computed using Equation 1. Comparisons of the $O\tilde{V}L$ were done using the standard bias and the relative bias. The standard bias is defined as the Monte Carlo mean minus the OVL, divided by the square root of the Monte Carlo variance. The relative bias is defined as the Monte Carlo mean minus the OVL, divided by the OVL. Comparisons of the bootstrap estimator of the variance of $O\tilde{V}L$ were made using the variance ratio and the relative bias of the variance. The variance ratio is defined as the ratio of the Monte Carlo variance to the bootstrap variance. The relative bias is defined as the bootstrap variance minus the Monte Carlo variance, divided by the Monte Carlo variance. The results of the Monte Carlo simulation are shown in the table in Appendix C.

Comparisons of the Monte Carlo mean to the OVL show that the estimator, $O\tilde{V}L$, is a biased estimator of OVL. This bias does not necessarily decrease as the sample size increases. The bias also is a function of the value of the OVL. When the two distributions were identical, the estimator tended to greatly under-estimate the value of the OVL in all of the distributional settings. The bias then decreased as the distributions became more distinct.

The bootstrap estimator of variance of the $O\tilde{V}L$ also performed well. The value of the variance decreased as sample size increased, thus suggesting that $O\tilde{V}L$ can be considered to be a consistent estimator of OVL. The relative bias of the estimator of the variance tended on average to be less than 10% for all distributional settings.

The relative bias of the $O\tilde{V}L$ for the sample size of 100 shows that for the normal case the bias was largest, 10.8%, when the two distributions were identical. Yet, this bias decreased to less than 2% when the two distributions became more distinct. As the value of the OVL decreased to less than 0.4000, the bias of the estimator slightly increased to just over 3%. When the two normal distributions were identical the $O\tilde{V}L$ greatly under-estimated the value of the OVL. For values of the OVL less than unity yet greater than .6000, the estimator tended to under-estimate the value of the OVL. The $O\tilde{V}L$ tended to over-estimate the value of the OVL for values of the OVL less than 0.600. Thus the estimator is a function of the value of the OVL. Also the standard bias was largest when the two distributions were identical. This bias decreased as the distributions became more dissimilar with the exception of the design points $\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 0, \sigma^2_2 = 3$ and $\mu_1 = 1, \sigma^2_1 = 1, \mu_2 = 0, \sigma^2_2 = 3$. The bootstrap estimator of variance performed well for this case as shown by the variance ratio being close to 1. As with the estimator of the OVL, the bootstrap estimator failed when the two distributions were identical. For this case the bootstrap estimator of the variance greatly overstated the apparent sampling variance of the kernel estimator of the OVL with the relative bias of the bootstrap estimator being greater than 50%. Yet, as the distributions became more distinct the relative bias was less than 10%. The bootstrap estimator tended to understate the apparent sampling variance of the $O\tilde{V}L$ with the exception of the design points $\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 1, \sigma^2_2 = 1$ and $\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 2, \sigma^2_2 = 4$.

For a sample size of 500 and two normal distributions, the bias was again largest when the two distributions were similar (i.e., approximately 5%). This bias decreased as the two distributions became more distinct with the relative bias being less than 7.5%. The bias of the estimator did not necessarily decrease with an increase in sample size. The standard bias was again largest for the case of equal distribution functions. Yet, this bias decreased dramatically as the two distributions became more distinct. The bias is a function of the value of the OVL with the estimator under-estimating the value of the OVL for values greater than .75 and over-estimating the value of the OVL for values less than .75. As with $n = 100$, the bootstrap estimator performed quite well (i.e., relative bias less than 7%) with the exception of the case of equal distribution functions. For this case, the bootstrap variance over-estimated the apparent sampling variance of the estimator with a relative bias of approximately 50%. For two dissimilar normal distributions, the bootstrap estimator of variance tended to under-estimate the value of the apparent sampling variance of the estimator with the exception of the design points $\mu_1 = 0, \sigma^2_1 = 3, \mu_2 = 2, \sigma^2_2 = 4$; and $\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 2, \sigma^2_2 = 4$.

For the Weibull distribution, the bias was largest, 10.5%, for $n_1 = n_2 = 100$ when the two distributions were identical. Yet, the bias decreased to less than 2% as the two distributions were distinct with the exception of the design points $\alpha_1 = 3.0, \beta_1 = 1.5, \alpha_2 = 1.5$ and $\beta_2 = 4.0$, where the bias was just over 5%. The $O\breve{V}L$ tended to under-estimate the true value of the OVL with the exception of the following design points: $\alpha_1 = 1.0, \beta_1 = 1.5$; $\alpha_2 = 1.0, \beta_2 = 3.0$; $\alpha_1 = 1.5, \beta_1 = 1.5$; and $\alpha_2 = 1.5, \beta_2 = 4.0$. The standard bias was largest for the case of identical distributions, yet it decreased as the distributions became more distinct. The bootstrap estimator of variance performed well with the exception of equal distributions. For this case the bootstrap variance over-estimated the apparent sampling variance of the estimator of the OVL with a relative bias of approximately 55%. The bias decreased as the distributions became more distinct (i.e., less than 10%), with the exception of α_1 = 1.0, β_1 = 2.0; α_2 = 1.0, β_2 = 3.5; α_1 = 3.0, β_1 = 1.5; and α_2 = 2.0, β_2 = 2.0, where the bias was approximate 14% and 11% respectively. The bootstrap variance tended to understate the apparent sampling variance of the estimator of the OVL when the two distributions differed.

For two Weibull distributions with the sample size of 500, the bias was again largest when the two distributions were similar, 4.7%. This bias decreased to less than 2.5% as the distributions were dissimilar with the exception of the design points α_1 = 1.5, β_1 = 1.5; α_2 = 2.0, β_2 = 2.0; α_1 = 3, β_1 = 1.5; and α_2 = 1.5, β_2 = 4.0. The bias did not necessarily decrease with the increase in sample size. The kernel estimator of the OVL tended to over-estimate the value of the OVL with the exception of the design points α_1 = 3.0, β_1 = 1.5; α_2 = 2.0, β_2 = 2.0; α_1 = 1.0, β_1 = 1.5; α_2 = 1.0, β_2 = 3.0; α_1 = 1.5, β_1 = 1.5; and α_2 = 1.5, β_2 = 4.0. Again the bootstrap estimator performed well (i.e., relative bias less than 8%) when the two distributions were dissimilar with the exception of the following design points: α_1 = 2.0, β_1 = 3.0 ; α_2 = 2.0, β_2 = 2.0 ; α_1 = 1.5 , β_1 = 4.0 ; and α_2 = 2.0, β_2 = 2.0 , where the bias was approximately 13% and 15.7%, respectively. The estimator failed when the two distributions were identical. For this case it over-estimated the apparent sampling variance of $O\tilde{V}L$, with the relative bias being just over 50%. The bootstrap variance tended to under-estimate the apparent sampling variance of the $O\tilde{V}L$ for all other cases studied with the exception of the design points α_1 = 1.5, β_1 = 1.5 and α_2 = 1.5, β_2 = 4.0.

For the two gamma distributions and sample size of 100 , the bias was largest, 10.6% when the two distributions were identical. The bias decreased to less than 4% as the distributions became more distinct. The standard bias was also largest for the case of equal distribution functions. The $O\tilde{V}L$ tended to under-estimate the value of the OVL for values of the OVL greater than .75. For OVL less than .75 the $O\tilde{V}L$ tended to over-estimate the true value of the OVL. The bootstrap variance estimator also performed well. The estimator failed for the case of identical distribution functions. In this case the bootstrap variance under-estimated the apparent sampling variance of the $O\tilde{V}L$, with a relative bias of over 50%. Yet, this bias decreased to approximately 7% as the two distributions became more distinct with the exception of the following design points: $\alpha_1 = 4.0$, $\alpha_2 = 3.5$; $\alpha_1 = 2.5$, $\alpha_2 = 3.5$; $\alpha_1 = 1.5$, $\alpha_2 = 2.5$ and $\alpha_1 = 2.5$, $\alpha_2 = 4.0$; where the bias was approximately 12% for these cases. The bootstrap estimator of the variance tended to understate the value of the apparent sampling variance of the estimator with the exception of the design points $\alpha_1 = 2.0$, $\alpha_2 = 4.0$; $\alpha_1 = 1.5$, $\alpha_2 = 3.5$; and $\alpha_1 = 1.5$, $\alpha_2 = 4.0$.

For two gamma distributions and sample size of 500, the bias was largest, approximately 5% when the two distributions were identical and for the design point $\alpha_1 = 1.5$, $\alpha_2 = 4.0$. The bias decreased to less than 3% for all other values of the OVL investigated. The bias did not necessarily decrease with an increase in sample size. The standard bias was largest when the two distributions were identical, yet decreased as the distributions became dissimilar. The kernel estimator of the OVL tended to over-estimate the value of the OVL with the exception of the design point $\alpha_1 = 4.0$, $\alpha_2 = 3.5$. The bootstrap estimator of variance performed well with the exception of the case of identical distributions. In this case, the bootstrap variance greatly overstated the apparent sampling variance of the $O\tilde{V}L$ with the bias being approximately 45%. As the distributions became more distinct, the bias decreased to less than 6% with the exception of the design point, $\alpha_1 = 2.5$, $\alpha_2 = 2.0$, where the bias was approximately 11%. For values of the OVL between .90 and .80, the bootstrap variance understated the apparent sampling variance of the $O\tilde{V}L$; for values of OVL between .75 and .65, the bootstrap variance overstated the apparent sampling variance of the $O\tilde{V}L$; the bootstrap variance understated the apparent sampling variance of the $O\tilde{V}L$ for values of the OVL between .65 and .50; and the bootstrap variance overstated the value of the apparent sampling variance of the $O\tilde{V}L$ for values less than .50.

The investigation of the four Beta distributional design points yield the following results. For the sample size of 100, the bias of the kernel estimator of the overlapping coefficient is largest, approximately 10.7%, when the two distributions are identical. The bias decreases to less than 3.5% as the two distributions became more distinct. The standard bias was also largest for identical beta distributions. The $O\tilde{V}L$ tended to under-estimate the value of the OVL for the design-points investigated. The bootstrap estimate of the variance of the $O\tilde{V}L$ performed well, with the exception of the case where the two beta distributions were identical. For the case of identical beta distributions, the bias was just over 42%, with the bootstrap estimate of variance greatly over-estimating the value of the apparent sampling variance of the $O\tilde{V}L$. When the two beta distributions differed the bias of the bootstrap variance decreased to less than 8% with the exception of the design point, $p_1 = 5$, $q_1 = 3$, $p_2 = 3$, and $q_2 = 3$, where the bias was approximately 15.5%.

For the sample size of 500, the relative bias of the kernel estimator of the overlapping coefficient was largest (4.7%) when sampling from two identical beta distributions. The bias decreased to less than 3% as the two distributions became more distinct. The bootstrap estimator of the variance performed well with the exception of identical distributions. When sampling from two identical beta distributions, the relative bias of the bootstrap estimator of variance was approximately 51%. In this case the bootstrap estimator tended to over-estimate the value of the apparent sampling variance of the $O\tilde{V}L$.

The relative bias of the bootstrap estimate decreased to less than 4% as the two distributions became more distinct with the exception of the design point $p_1 = 2$, $q_1 = 2$, $p_2 = 3$, and $q_2 = 3$, where the bias was just over 10%. When the distributions differed, the bootstrap estimator tended to under-estimate the value of the apparent sampling variance of the $O\tilde{V}L$.

The $O\tilde{V}L$ was also explored for mixtures of distributions. For the standard normal and standard Cauchy distributions, $O\tilde{V}L$ under-estimated the value of the OVL for the sample size of 100 with the bias being 2.2%, while $O\tilde{V}L$ over-estimated the value of the OVL for the sample size of 500 with the bias being less than 1%. The bootstrap estimate of the variance performed well. For the sample size of 100, the bootstrap estimator of the variance under-estimated the apparent sampling variance of the estimator with the bias being 4.5%. The bootstrap estimator also understated the apparent sampling variance of the estimator by 3.2% for the sample size of 500.

For a gamma distribution with $\alpha = 3$ and a chi-squared distribution with 4 degrees of freedom, the $O\tilde{V}L$ under-estimated the value of the OVL for both sets of sample sizes with this bias being 4.2% for $n = 100$ and 2.5% for $n = 500$. The bootstrap estimate of the variance understated the apparent sampling variance of the estimator with the bias being 11% for n = 100 and 3.5% for $n = 500$. In both cases, the bootstrap estimate of the variance tended to understate the value of the apparent sampling variance of the estimator.

Comparison of the Kernel Estimator of the OVL to the Maximum Likelihood Estimator of the OVL

The maximum likelihood estimator of the OVL was computed for 1,000 simulations using the nine design points used in the investigation of the kernel estimator for the OVL for sample of $n_1 = n_2 = 100$ and 500, where the variance of the two normal distributions differed (see Appendix D for SAS program used for the simulation study). The maximum likelihood estimator for the reparameterized overlapping coefficient developed by Clemons (1996) was used to examine the accuracy of the kernel estimator of th OVL when sampling from two normal distributions with unequal variances. The Monte Carlo mean and variance were computed from the first and second moments from the 1,000 simulations. Also, the standard bias and relative bias were computed as described previously. To compare the bias of the kernel estimator to that of the maximum likelihood estimator, the standard bias of the kernel estimator was calculated as follows: the difference of the Monte Carlo mean minus OVL divided by the Monte Carlo standard error of maximum likelihood estimator (see table in Appendix D). In units of the standard error of maximum likelihood estimator of the OVL, it is shown that the bias of the kernel estimator is greater than the bias of the maximum ikelihood estimator with the exception of the following design points: $\mu_1 = 0$, $\sigma^2{}_1 = 1$, $\mu_2 =$ 0 and $\sigma^2{}_2 = 3$, $n_1 = n_2 = 500$; $\mu_1 = 5$, $\sigma^2{}_1 = 10$, $\mu_2 = 3$; and $\sigma^2{}_2 = 5$, $n_1 = n_2 = 100$. It must be noted that although the bias of the kernel estimator of the OVL was greater than the maximum likelihood estimator, this difference in most cases was not drastic.

The relative inefficiency of the kernel estimator of the OVL compared to the maximum likelihood estimator of the OVL as estimators of the OVL between two normal distributions is indicated by the ratio of their Monte Carlo variances, also shown in Appendix

F. The variance of the kernel estimator is approximately 1.24 times the variance of maximum-likelihood estimator, running from low of 1.00466 ($\mu_1 = 0$, $\sigma^2_1 = 3$, $\mu_2 = 2$ and $\sigma^2_2 = 4$, $n_1 = n_2 = 500$) to a high of 1.61661 ($\mu_1 = 1$, $\sigma^2_1 = 1$, $\mu_2 = 2$ and $\sigma^2_2 = 4$, $n_1 = n_2 = 500$). The ratio of the Monte Carlo variances of the kernel estimator and the maximum-likelihood estimator increases with sample size when the value of the OVL is less than 65% and decreases when the value of the OVL is greater than 65%. This shows that the relative inefficiency of the kernel estimator of the OVL is a function of the value of the OVL. Thus when using the kernel estimator of the OVL when the samples follow a normal distribution, the estimator is on average 80% efficient.

Modeling of the Bias

A regression analysis (Montgomery, 1991) was conducted to fit models for predicting the OVL given the value of the kernel estimator of the OVL and sample size. These models were developed to reduce the bias of the estimator of the overlapping coefficient for the distributions and design points used in the Monte Carlo simulation study.

For the normal distribution it was noticed that for large values of the overlapping coefficient (i.e. OVL < 0.6000) the kernel estimator of the OVL under-estimated the OVL; the kernel estimator tended to over-estimate the value of the OVL for values of the OVL less than 0.600. The following cubic model was fit to the data:

$$\text{OVL} = 0.208039 - 0.00005429{*}N + 1.426574{*}\text{MCOVL}^2 - 0.530713{*}\text{MCOVL}^3 \quad (37)$$

Using this model for larger values of the OVL, when the kernel estimator under-estimates the value of the OVL, the model tends to increase the estimate of the OVL. For smaller values of the OVL, when the kernel estimator overstates the value of the OVL, the model tends to decrease the estimate of the OVL. Also the model adjusts by slightly reducing the estimator of the OVL with an increase in sample size. This model adjusts the bias of the kernel estimator of the overlapping coefficient assuming that the data from the two samples come from a normal distribution and also that the two samples are of equal sample size. Appendix G shows the percent change in the relative bias using the model. The model works reasonably well in reducing the bias of the kernel estimator of the overlapping coefficient.

For the Weibull distributions the kernel estimator tended to overstate the value of the OVL with the bias being largest for larger values of the OVL (i.e. OVL = 1). Thus, a quadratic model was fit to the data. The model is as follows:

$$\text{OVL} = 0.333109 - 0.0000712874 * N + 0.779401 * \text{MCOVL}^2 \qquad (38)$$

This model adjusts the kernel estimator by increasing the estimator of the overlapping coefficient thus reducing the bias. As the value of the OVL increases, the model adjusts for the increase in bias by increasing the value of the estimator. The model adjusts for the increase in sample size. The change in relative bias given in Appendix E shows that the model performs reasonably well in reducing the bias of the kernel estimator when sampling from two Weibull distributions with equal sample size.

When sampling from two Gamma distributions, the kernel estimator tended to overstate the value of the OVL for the smaller values of the OVL (< 0.7000) and under-estimate the value of the OVL for the larger values of the OVL. The best model for the data was as follows:

$$\text{OVL} = 0.296354 - 0.00006970 * N + 0.832176 * \text{MCOVL}^2 . \qquad (39)$$

With this model, for smaller values of the overlapping coefficient when the kernel estimator tends to overstate the value of the OVL, the estimator is reduced. For larger values where the kernel estimator tends to understate the value of the OVL, the estimator is increased. The model also adjusts for the increase in sample size. Appendix E again shows that the model for reducing the bias of the kernel estimator when sampling from two gamma distributions with equal sample size, performs reasonably well.

Appendix G contains the analysis of variance (ANOVA) tables and the R-squared values; tests of the various parameters; and the predicted values produced given the Monte Carlo estimates of the OVL, relative biases, and percent change of the bias of the estimator for the above models. Again, these models performed quite well, with the bias being significantly reduced for the case of sampling from two identical distributions with equal sample sizes.

III. ALTERNATIVE NORMAL REFERENCE RULE FOR THE BANDWIDTH OF THE NAIVE KERNEL ESTIMATOR

As noted previously, the kernel estimator using the normal reference rule did not asymptotically reduce the value of the bias of the $O\tilde{V}L$. We know from the previous discussion of the kernel density estimator that the estimator is dependent on the value of the shaping parameter or bandwidth, h. As the sample size increases, the value of the shaping parameter is reduced thus increasing the number of jump points in the distribution. While this is desirable when one is using the kernel technique in estimating the density, it may also have an adverse effect on estimation the value of the OVL. This can be alleviated by increasing the value of the shaping parameter. In this chapter we will explore the kernel density estimator using an alternative value of the shaping parameter.

Rosenblatt (1956) showed that for the integrated expected mean-square error criterion, the best shaping parameter, h, was a constant times $n^{-1/5}$. This technique was also used by Scott (1992) in development of the normal reference rule. Waterman and Whiteman (1978) suggested an alternative method of determining an optimum value of the shaping parameter. By using the Kolmogrov-Smirnov statistic, they were able to obtain bounds for the shaping parameter using only properties of the first derivative of the density function as follows. It must be briefly noted that Rosenblatt (1956) required the existence of three derivatives of the density function.

The process for determing the optimal bandwidth using the Kolmogrov-Smirnov statistic is as follows. We have that

$$\left|\frac{Y(x)}{2nh}-\frac{F(x+h)-F(x-h)}{2h}\right| \leq \left|\frac{Y(x)}{2nh}-\frac{F(x+h)-F(x-h)}{2h}\right| + \left|\frac{F(x+h)-F(x-h)}{2h}-f(x)\right| , \tag{40}$$

where,

$$Y(x) = n(F_n(x+h)-F_n(x-h)) . \tag{41}$$

Then to bound the first quantity on the right-hand side,

$$\left|\frac{Y(x)}{2nh}-\frac{F(x+h)-F(x-h)}{2h}\right| \leq \left|\frac{F_n(x+h)-F_n(x-h)}{2h}\right| + \left|\frac{F_n(x+h)-F_n(x-h)}{2h}\right| . \tag{42}$$

If we let $D_n(\alpha)$ satisfy the following

$$P(\max_x |F_n(x)-F(x)| > \frac{D_n(\alpha)}{h} = \alpha, \tag{43}$$

then

$$\left|\frac{Y(x)}{2nh}-\frac{F(x+h)-F(x-h)}{2h}\right| \leq \frac{D_n(\alpha)}{h} \tag{44}$$

with probability of at least $1 - \alpha$. To bound the second quantity of the right hand side, it should be noted that

$$F(x \pm h) = F(x)+F'(x)(\pm h)+\frac{F''(x_i)(\pm h^2)}{2} . \tag{45}$$

Thus,

$$\left|\frac{F(x+h)-F(x-h)}{2h}-f(x)\right|=|f(x)+\frac{h}{4}(f'(x_1)-f'(x_2))-f(x)|=\frac{h}{4}|f'(x_1)-f'(x_2)|\,. \qquad (46)$$

If $|f'(x)| \leq C$ for all x, then

$$\left|\frac{F_n(x+h)-F_n(x-h)}{2h}-f(x)\right|\leq\frac{D_n(\alpha)}{h}+\frac{Ch}{2}\doteq B_1(h)\,. \qquad (47)$$

The h which minimizes $B_1(h)$ is

$$h=(2D_n(\alpha)/C)^{1/2}\,. \qquad (48)$$

If we let $D_n(\alpha) = K(\alpha)/n^{-1/2}$, the asymptotic form given in Lindgren, 1968, then

$$h=(2K(\alpha)/C)^{1/2}n^{-1/4}\,. \qquad (49)$$

The results hold uniformly for all x with probability at least $1 - \alpha$ (Waterman & Whiteman, 1978). Thus using the above shaping parameter an alternative normal reference rule can be obtained by using the normal density function for the value of C. From the normal density we obtain $|f(x)| \leq(\sqrt{(2\pi)}e\sigma^2)^{-1}$. If we let $\alpha = 0.05$, then $K(\alpha) = 1.36$ and we obtain a shaping parameter of $h = 3.35\sigma n^{-1/4}$. If σ is unknown we can use the usual sample standard deviation, s, as an estimator of σ. The kernel density estimator, $O\tilde{V}L$, can then be estimated as before yet incorporating the alternative value of the shaping parameter.

Let us consider for example the two kernel estimated densities in Figures17-19 which are obtained from two samples of size 500 generated from two normal distributions. The first sample is generated from the standard Normal distribution; the density estimate derived from this sample is indicated by a dotted line. The second is from a normal distribution with mean = 1 and variance = 1. Using the subroutines described in Appendix A, we find that the $O\tilde{V}L$ is 0.688772. The actual overlap is 0.617075.

Monte Carlo Investigation

A Monte Carlo simulation study was used to investigate the properties of the kernel estimator of the OVL using the alternative normal reference rule. Using the design points used in the previous simulation study for two normal distributions and sample sizes of 100 and 500, $O\tilde{V}L$ was computed. The FORTRAN programs described in chapter II and shown in Appendix A were used for the study with an adjustment made to the value of the shaping parameter. The results of the Monte Carlo study are summarized in Appendix H.

Discussion

The kernel estimator of the overlapping coefficient using the alternative rule is again a biased estimator. When the two distributions were identical, use of the alternative reference rule results in a bias that is smaller than the bias of the kernel estimator of the OVL using the normal reference rule discussed in chapter II. Yet, when the two distributions became more distinct, the bias dramatically increased with the exception of OVL = 0.740641 (i.e., $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu = 0, {}_2^2\sigma = 3$). As with the normal reference rule, the bias of the $O\tilde{V}L$ calculated using the alternative reference rule did not necessarily decrease with an increase in sample size. Thus the use of the this alternative reference rule, which

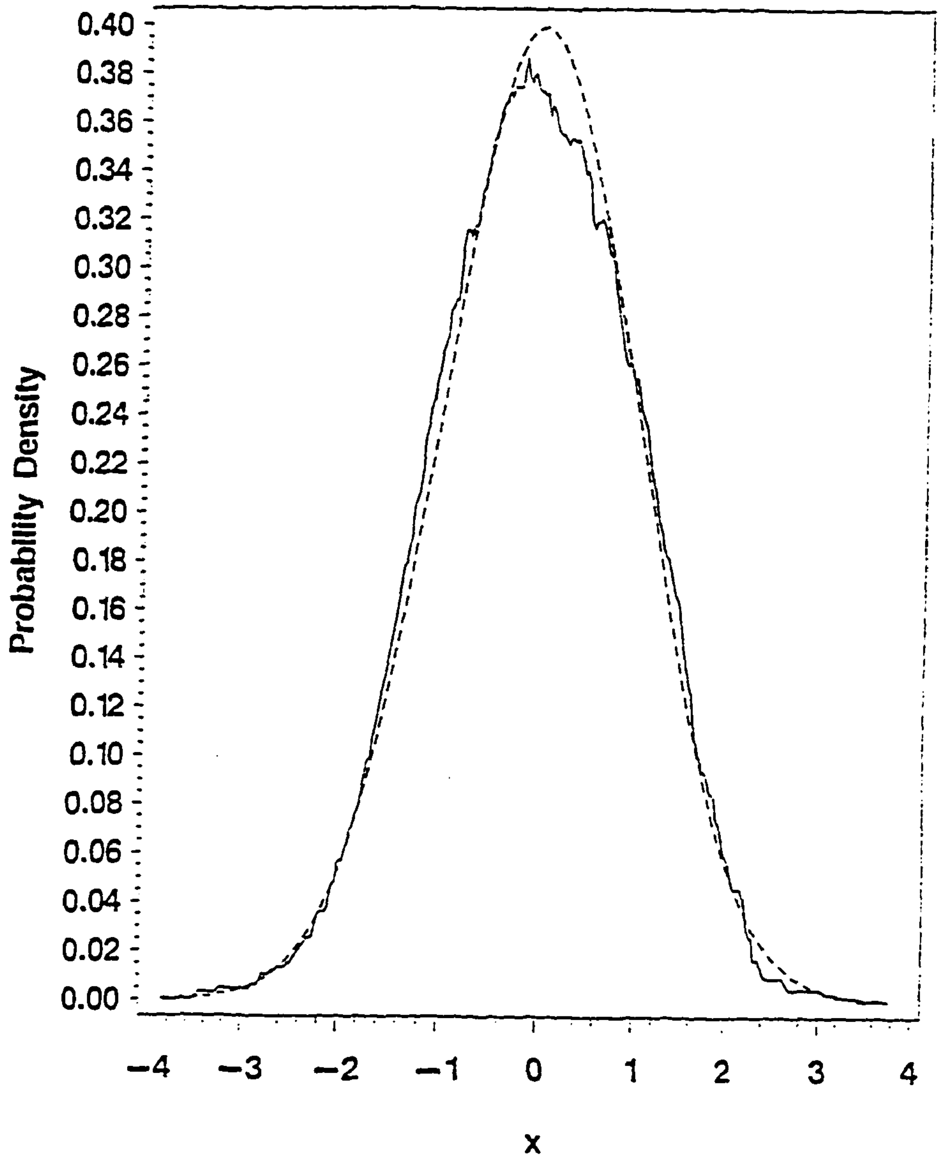

Figure 17. The kernel density estimator for a standard normal distribution using the alternative rule. The solid line denotes a kernel density estimator generated from pseudo normal random deviates. The dotted line denotes a standard normal distribution

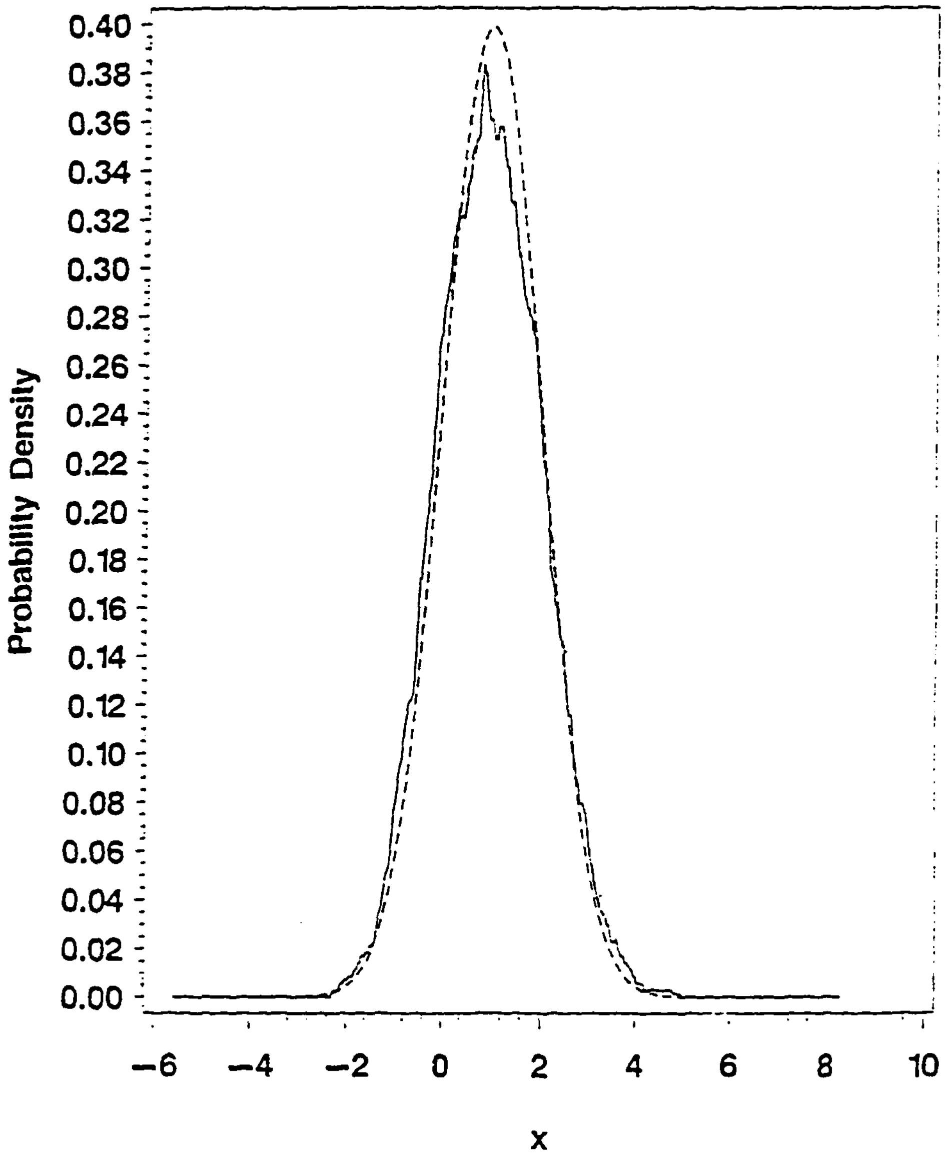

Figure 18. The kernel density estimator for a normal distribution using the alternative rule. The solid line denotes a kernel density estimator generated from pseudo normal random deviates. The dotted line denotes a normal distribution with mean = 1 and variance = 1.

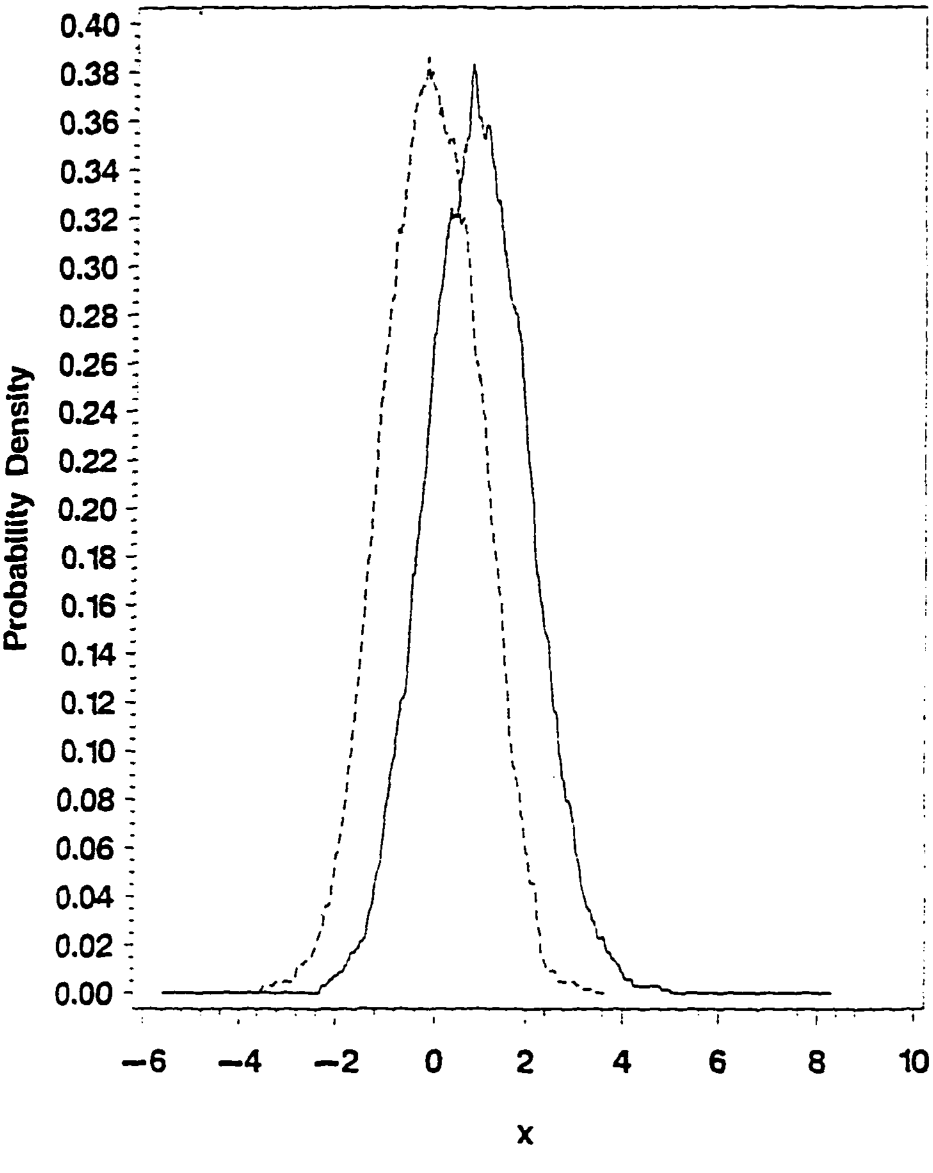

Figure 19. The overlap using the kernel densities for two normal distributions using the alternative rule. The solid line denotes a kernel density estimator generated from pseudo normal random deviates. The dotted line denotes a kernel density estimator generated from pseudo standard normal random variates.

increases the shaping parameter used to compute the kernel estimator of the OVL, only improves the kernel estimator when the two distributions are identical. For distinct distributions, the alternative rule does not improve the estimator of the OVL. Since the objective of this study is to develop a kernel density estimator which is robust for all distributional settings, sample size, and values of the overlap between the two distributions, it is recommended that the normal reference rule used for the development of the kernel estimator of the OVL in chapter II be employed.

IV. EXAMPLES OF THE USES OF THE OVL

Selectivity of the Migration of Farmers Between 1850 and 1860

As an example of the use of $O\tilde{V}L$, let us consider one part of a study designed to investigate the selectivity of the migration of Alabama farmers between 1850 and 1860 (Inman, 1981) which was used by Inman (1984) as an example of the OVL for two normal cases with equal variances. A simple random sample of 664 farm operators was obtained from the 1850 census of agriculture for ten Alabama counties. Each farm operator in the sample was matched to the corresponding entries for his household and his slave-force in the 1850 census of free population and slave population; from this information his wealth in 1850 was estimated. Those farm operators in the sample who could be located in the same county in the 1860 census are classified as persistent farmers. Those who were not found in the 1860 census of the county in which they resided in 1850 did not persist. (A rudimentary adjustment for the effect of mortality, not discussed here, is also made.) We shall concern ourselves with a subset of this sample, consisting of 601 male farm operators who were listed as the heads of their households in the census of free population and for whom consistent census data is available.

Since the distribution of the data was highly skewed, Inman used a logarithmic transformation in his evaluation of the $O\hat{V}L$. This same transformation will be used in this examination of $O\tilde{V}L$ (see Appendix I) for comparison to Inman's cubic spline estimator of the OVL. A test for normality was done for each sample using the Shapiro-Wilks statistic

(Shapiro & Wilks, 1974). For the 317 persistent farmers, the *p* value for the test was 0.0001 and for the 284 non persistent farmers the *p* value was 0.0266. Thus we can reject the null hypothesis in both cases that the two distributions are normal.

Using the natural logarithms, the sample median for the persistent farmers is 7.34225. For the nonpersistent farmers, the sample median is 6.80445. A nonparametric test, the Median Two Sample Test using a normal approximation, yields a z value of -3.22691 which is statistically significant at the 0.0013 level. Thus it appears reasonable to assume that the median wealth of persistent Alabama farmers exceeds the median wealth of the nonpersistent counterparts, indicating that the migration of Alabama farm operators between 1850 and 1860 to some degree selected poorer farmers.

According to Inman (1984), the degree of selectivity depends not on the difference in population medians but instead on the actual difference in the distribution of wealth of the two groups of farmers. If the distributions are highly distinct, then a strong case can be made for migration selective with respect to wealth. The wealthy farmers were able to persist while the poorer farmers were forced to relocate. The $O\tilde{V}L$ obtained for the two groups was 0.87771356, which indicates that the distributions of wealth for these two groups of Alabama farmers are not as distinct as a simple comparison of the sample medians might suggest (see Figures 20-23) . Therefore, we can conclude that the difference in wealth for the farmers who persisted and those who did not persist is not as distinct as suggested by a comparison of the medians.

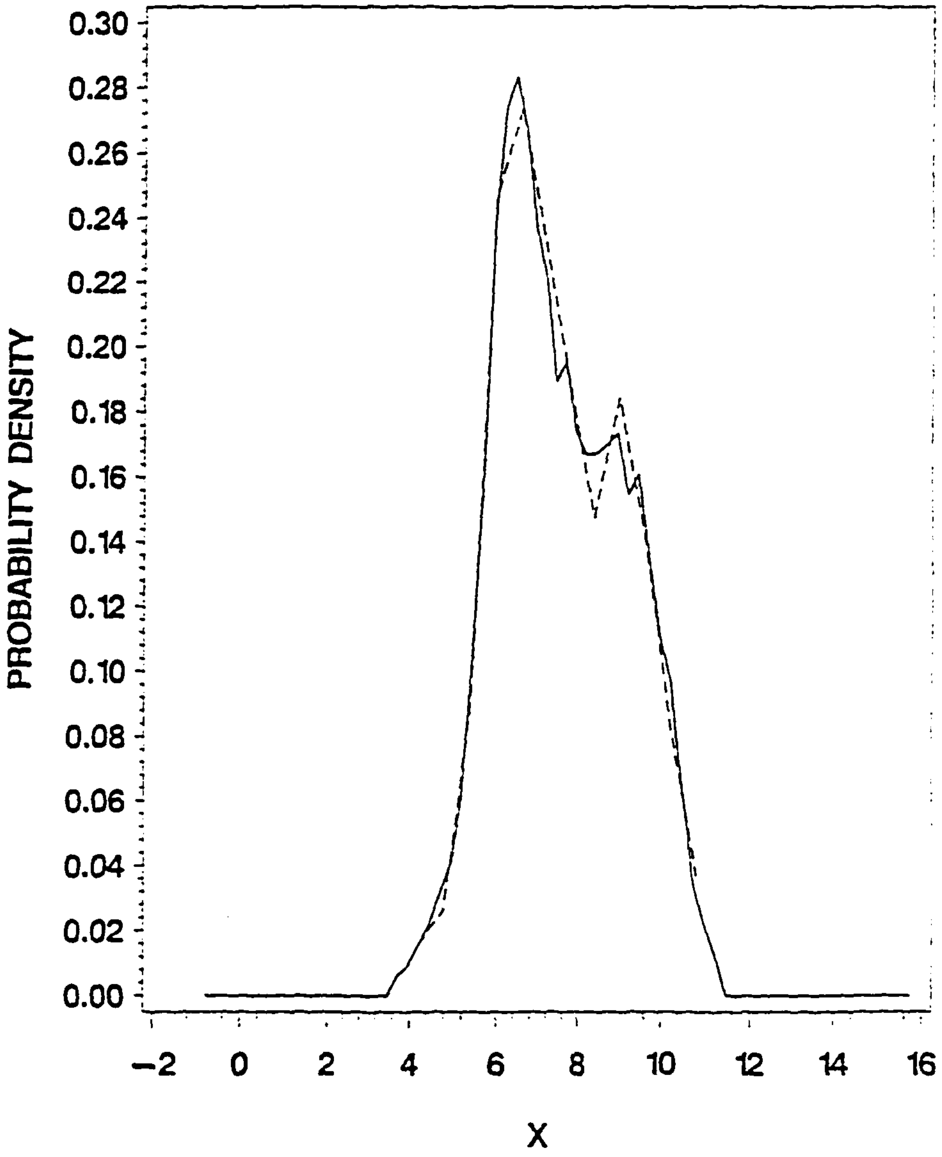

Figure 20. The kernel density estimator for the wealth of the farmers who persisted to 1860. The dotted line denotes the midpoints of the histogram of the data. The solid line denotes the kernel density estimator for the data.

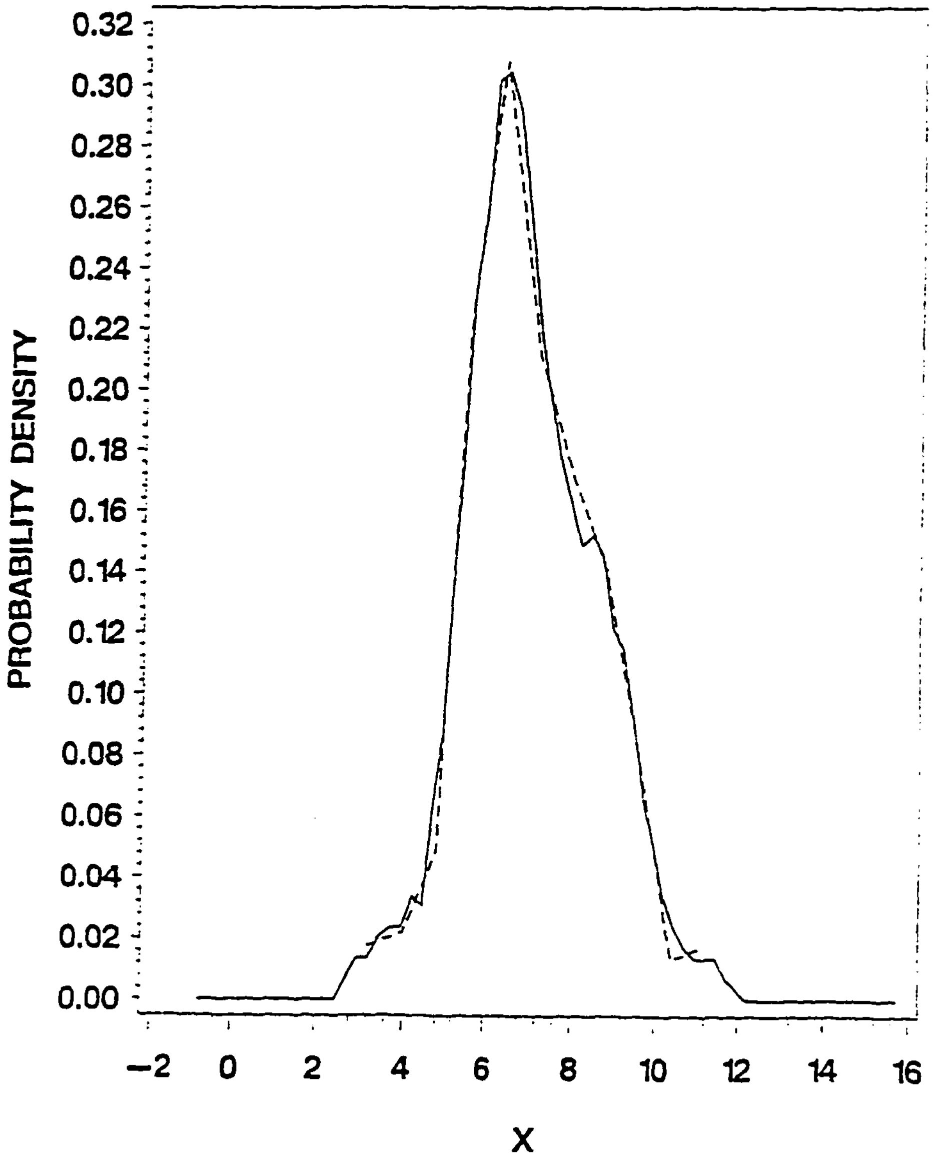

Figure 21. The kernel density estimator for the wealth of the farmers who did not persist to 1860. The dotted line denotes the midpoints of the histogram of the data. The solid line denotes the kernel density estimator for the data.

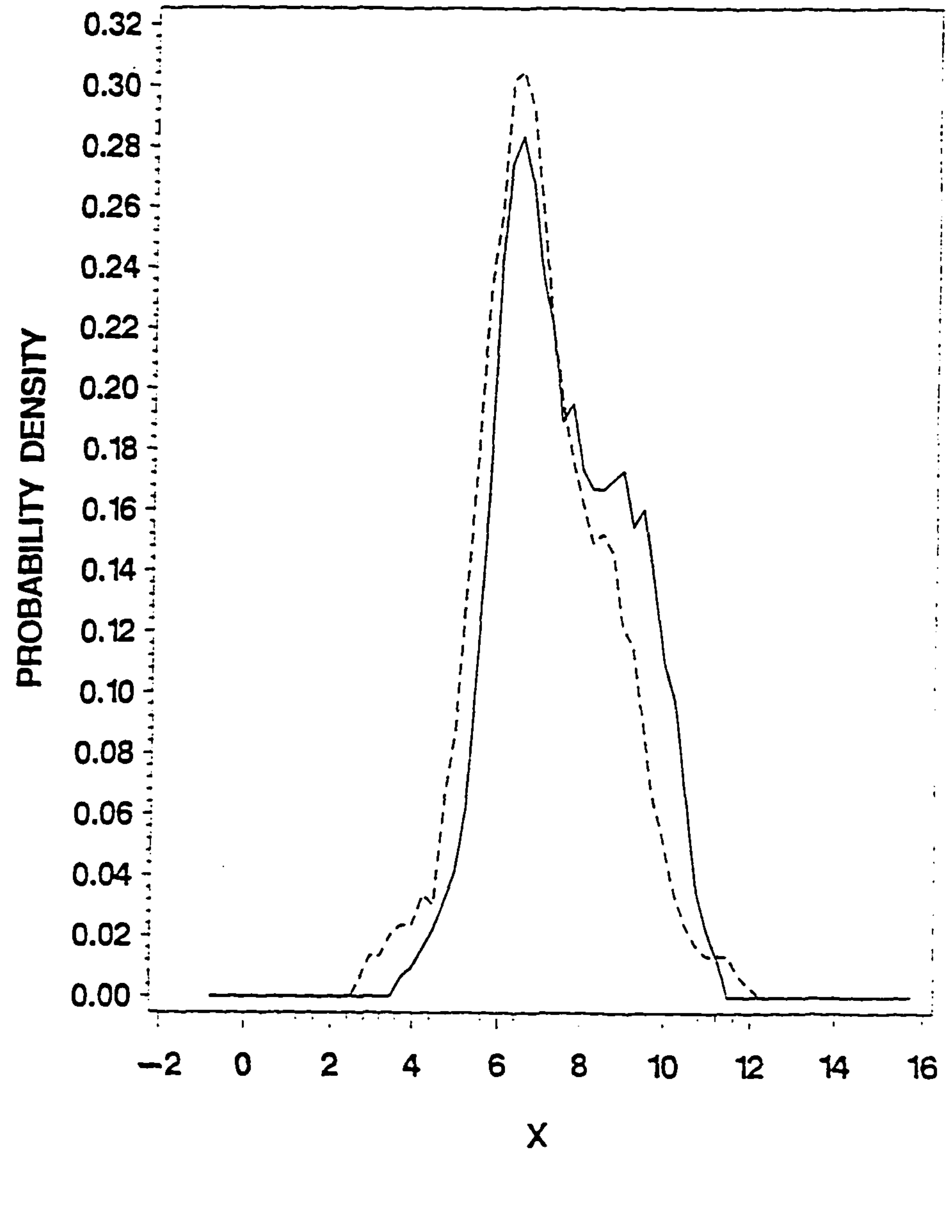

Figure 22. The kernel estimator of the overlap for the wealth data. The solid line denotes the kernel density estimator for farmers who persisted to 1860. The dotted line denotes the kernel density for farmers who did not persist to 1860.

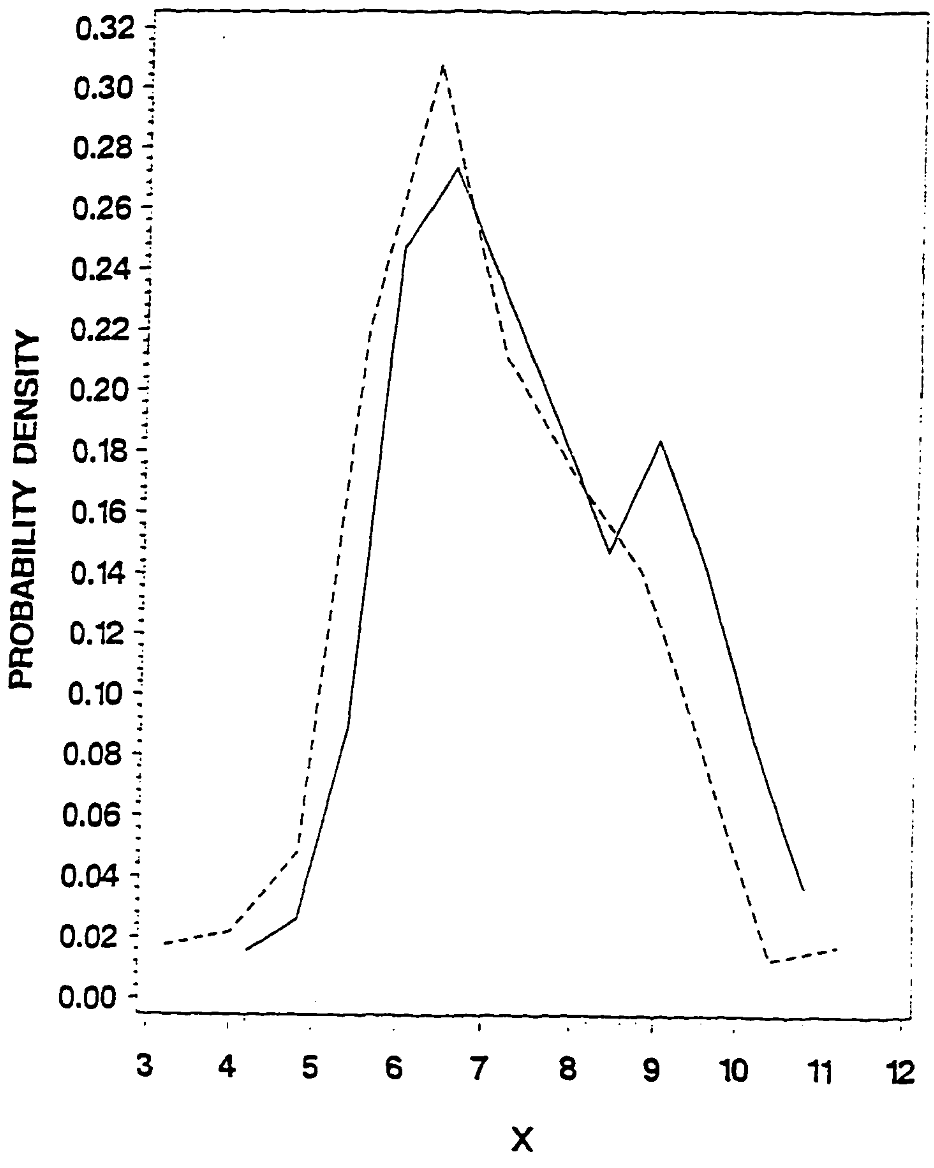

Figure 23. The overlap for the wealth data. The solid line denotes the midpoints for the histogram for farmer who persisted to 1860. The dotted line denotes the midpoints of the histogram for farmers who did not persist to 1860.

Bootstrap estimates of the variance of $O\tilde{V}L$ were obtained using Equation 34. The results for three different values of B are as follows:

$$\text{For B} = 100, \quad \tilde{Var}_B(O\tilde{V}L) = 0.000919;$$

$$\text{For B} = 200, \quad \tilde{Var}_B(O\tilde{V}L) = 0.000878;$$

$$\text{For B} = 500, \quad \tilde{Var}_B(O\tilde{V}L) = 0.000823.$$

Using the results obtained when B = 500, the percentile method for constructing a bootstrap confidence interval for OVL described previously was used to compute a confidence interval for OVL using the 1850 wealth data. A 90% confidence interval for the true overlap between wealth distribution of the persistent and nonpersistent Alabama farmers, using the bootstrap distribution constructed from the 500 $O\tilde{V}L^{*}$, is given by

$$[F^{*\,-1}_{500}(0.05), F^{*\,-1}_{500}(0.95)] = (0.816950\,, 0.901633).$$

Figure 24 is a histogram of the 500 bootstrap estimates. A Shapiro-Wilks test for normality was performed to test whether the empirical distribution of the bootstrap estimates can be considered normal. The *p* value for the test was 0.1242.

Inman (1984) considered this example in his work with the maximum likelihood estimator and the cubic spline estimators of overlapping coefficient. Using the maximum likelihood estimator, the value of the estimator was 0.859614 with a 90% confidence interval of (0.808967 , 0.915465). For the cubic spline estimator of the OVL, the value of the estimator was 0.869152 with a 90% confidence interval of (0.848472, 0.941238) . Our value of the kernel estimator of the OVL, 0.877714, was larger than both values computed in Inman (1984).

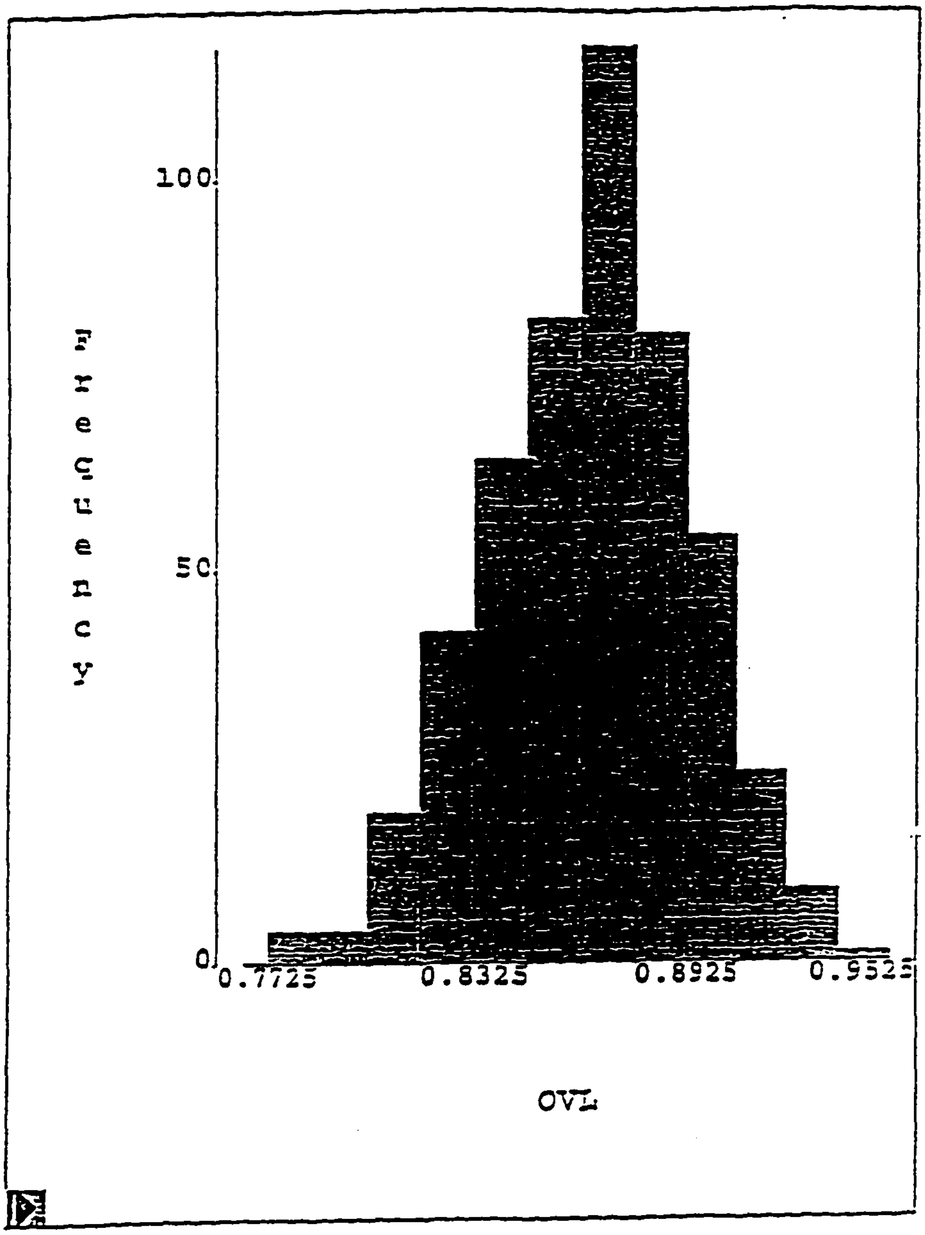

Figure 24. The histogram of the B = 500 bootstrap estimators for the wealth data.

Irish Education Transition Data

Next we considered data on the Irish Education Transition for a sample of 469 Irish school children aged 11 in 1976 (Greaney and Kelleghan, 1984). Each student was classified by sex and the measure of interest was the students' Drumcondra Verbal Reasoning Test score (see Appendix I). A test for normality using the Shapiro-Wilks test for the 231 males yields a *p* value of 0.011. While the test for normality for the 238 females yields a *p* value of 0.2783. Since the male sample cannot be considered to follow a normal distribution, a nonparametric test was conducted to test differences in the median test scores between males and females. The median Drumcondra Verbal Reasoning Test score for males is 104. While the median Drumcondra Verbal Reasoning Test score for females is 100.5. A Median Two Sample Test using a normal approximation yields a z statistic of 2.04991, which is statistically significant at the 0.0404 level. Thus it appears reasonable to assume that the median Drumcondra Verbal Reasoning Test score for males exceeds the median Drumcondra Verbal Reasoning Test score for females. The $O\tilde{V}L$ for the two distributions was 0.85528, which suggests that the distributions may not be as distinct as suggest by the simple comparison of medians (see Figures 25-28). A re-evaluation of the data shows that the Drumcondra Verbal Reasoning Test scores for female students tended to be concentrated around the median test score. While for male students, scores tended to be on the higher end of the distribution thus causing the distribution to be right skewed. Thus the difference is more in the right tails of the distributions, where elsewhere the two distributions overlapped.

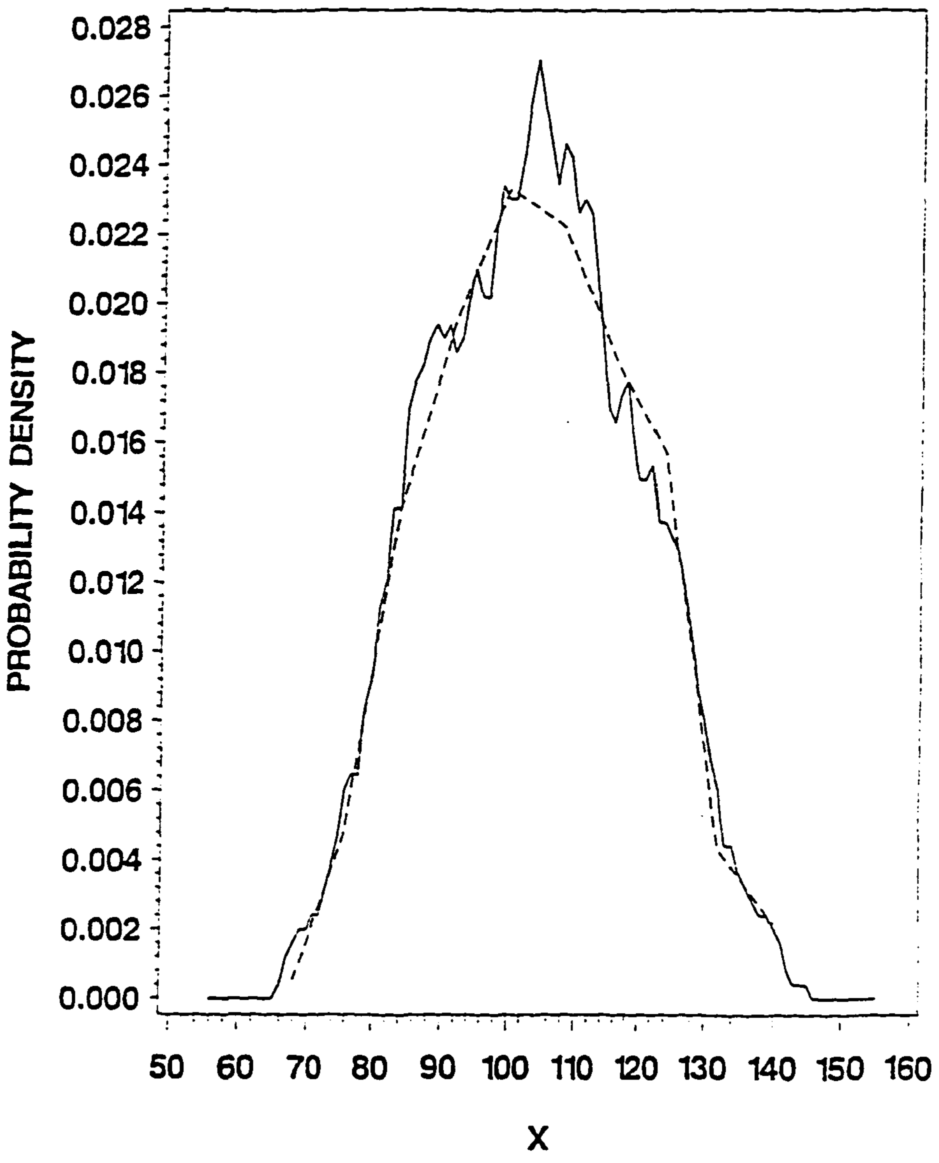

Figure 25. The kernel density estimator for the Drumcondra Verbal Reasoning Test scores for the male students. The dotted line denotes the midpoints of the histogram of the data. The solid line denotes the kernel density estimator for the data.

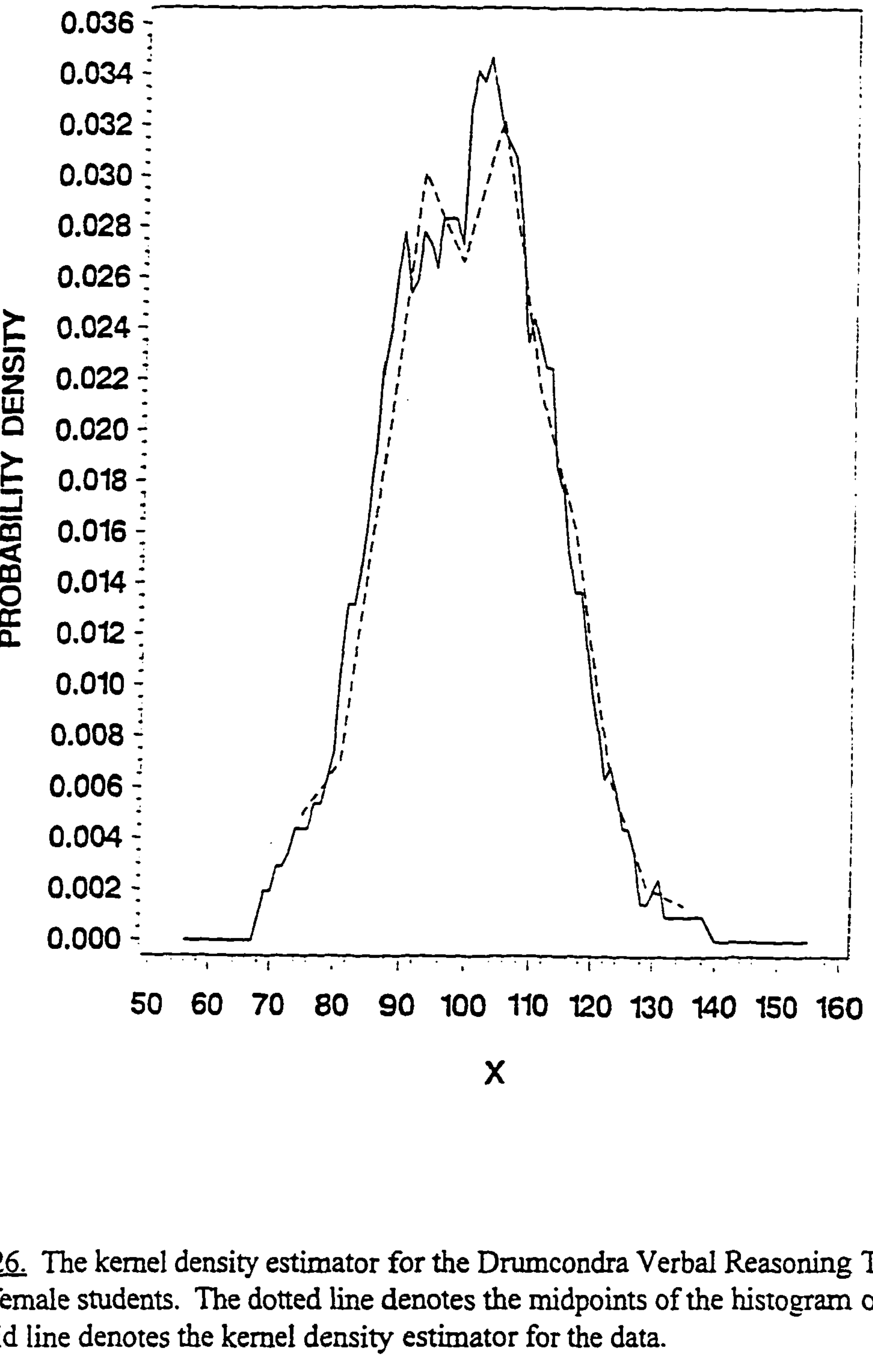

Figure 26. The kernel density estimator for the Drumcondra Verbal Reasoning Test scores for the female students. The dotted line denotes the midpoints of the histogram of the data. The solid line denotes the kernel density estimator for the data.

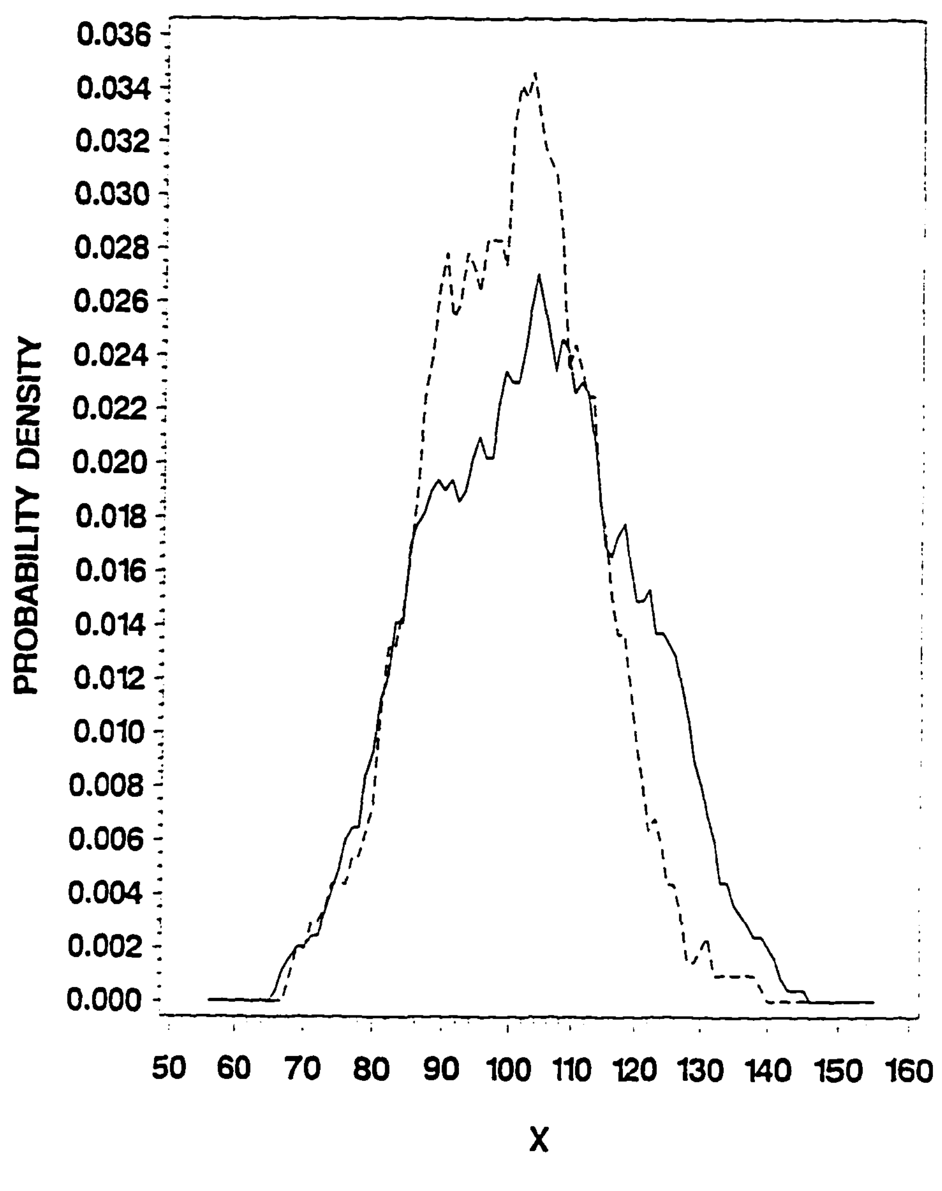

Figure 27. The kernel estimator of the overlap for the Irish Education Transition data. The solid line denotes the kernel density estimator for the scores of male students. The dotted line denotes the kernel density for the scores of the female students.

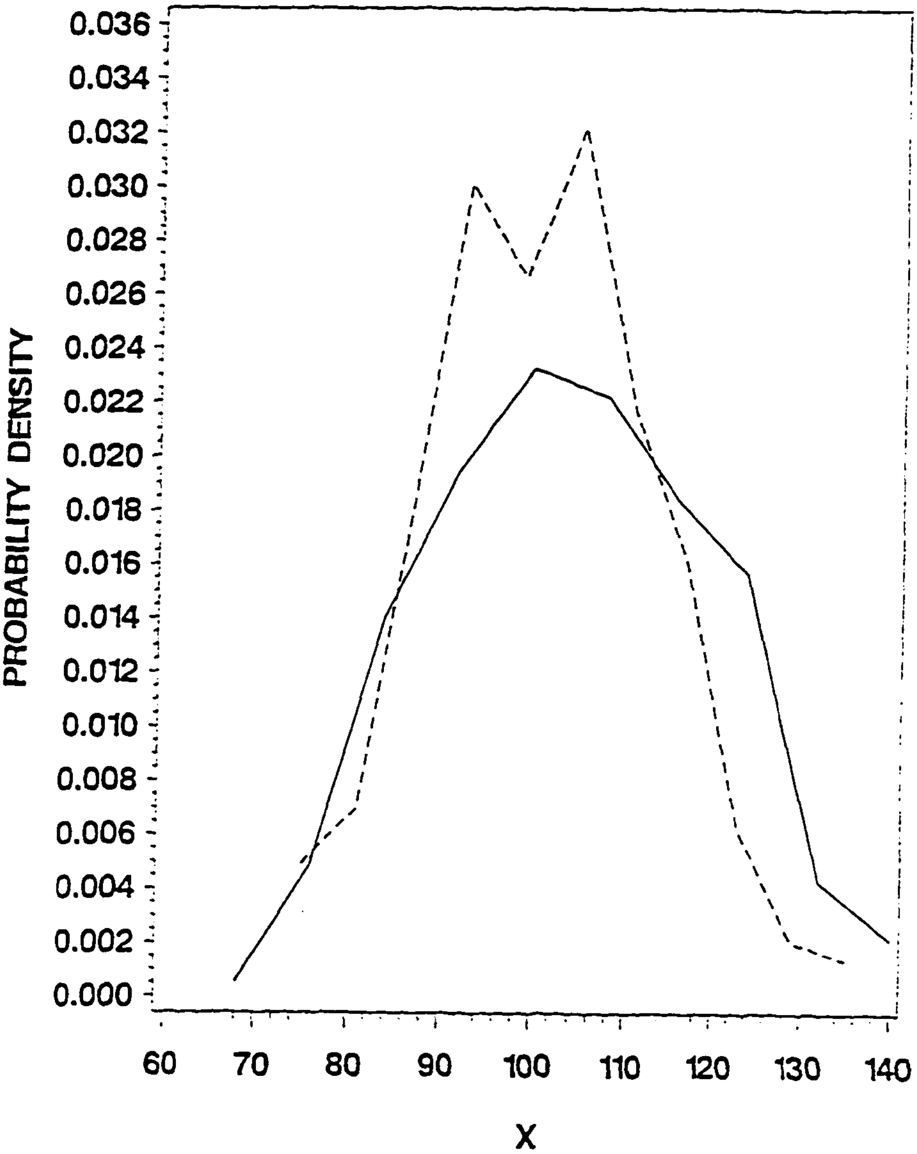

Figure 28. The overlap for the Irish Education Transition data. The solid line denotes the midpoints for the histogram for the scores for male students. The dotted line denotes the midpoints of the histogram for the scores for female students.

Bootstrap estimates of the variance of $O\tilde{V}L$ were obtained using Equation 34. Results for three different values of B are as follows:

$$\text{For B} = 100,\quad \tilde{Var}_B(O\tilde{V}L) = 0.000874;$$

$$\text{For B} = 200,\quad \tilde{Var}_B(O\tilde{V}L) = 0.000929;$$

$$\text{For B} = 500,\quad \tilde{Var}_B(O\tilde{V}L) = 0.000823.$$

Using the results obtained when B = 500, the percentile method for constructing a bootstrap confidence interval for OVL described previously was used to compute a confidence interval for OVL using the 469 test scores. A 90% confidence interval for the true overlap between Drumcondra Verbal Reasoning Test scores for females versus male students, using the bootstrap distribution constructed from the 500 $O\tilde{V}L^*$, is given by

$$[F^{*\,-1}_{500}(0.05), F^{*\,-1}_{500}(0.95)] = (0.788187, 0.891234).$$

Figure 29 is a histogram of the 500 bootstrap estimates. A Shapiro-Wilks test for normality was performed to test whether the empirical distribution of the bootstrap estimates can be considered normal. The *p* value for the test was 0.0731 .

Acute Myocardial Infarction Registry

For a last example of the use of $O\tilde{V}L$, we considered data from the Acute Myocardia Infarction Registry (Rogers, Dean, Moor, Wool, Burgard, & Bradley, 1993). A simple random sample of 1,156 patients were obtained from the registry. Each patient was identified by two variables: gender and whether or not the patient experienced chest pain for more than 6 hours before treatment. The response variable of interest was minutes from onset of ischemic chest pain to ECG (see Appendix I).

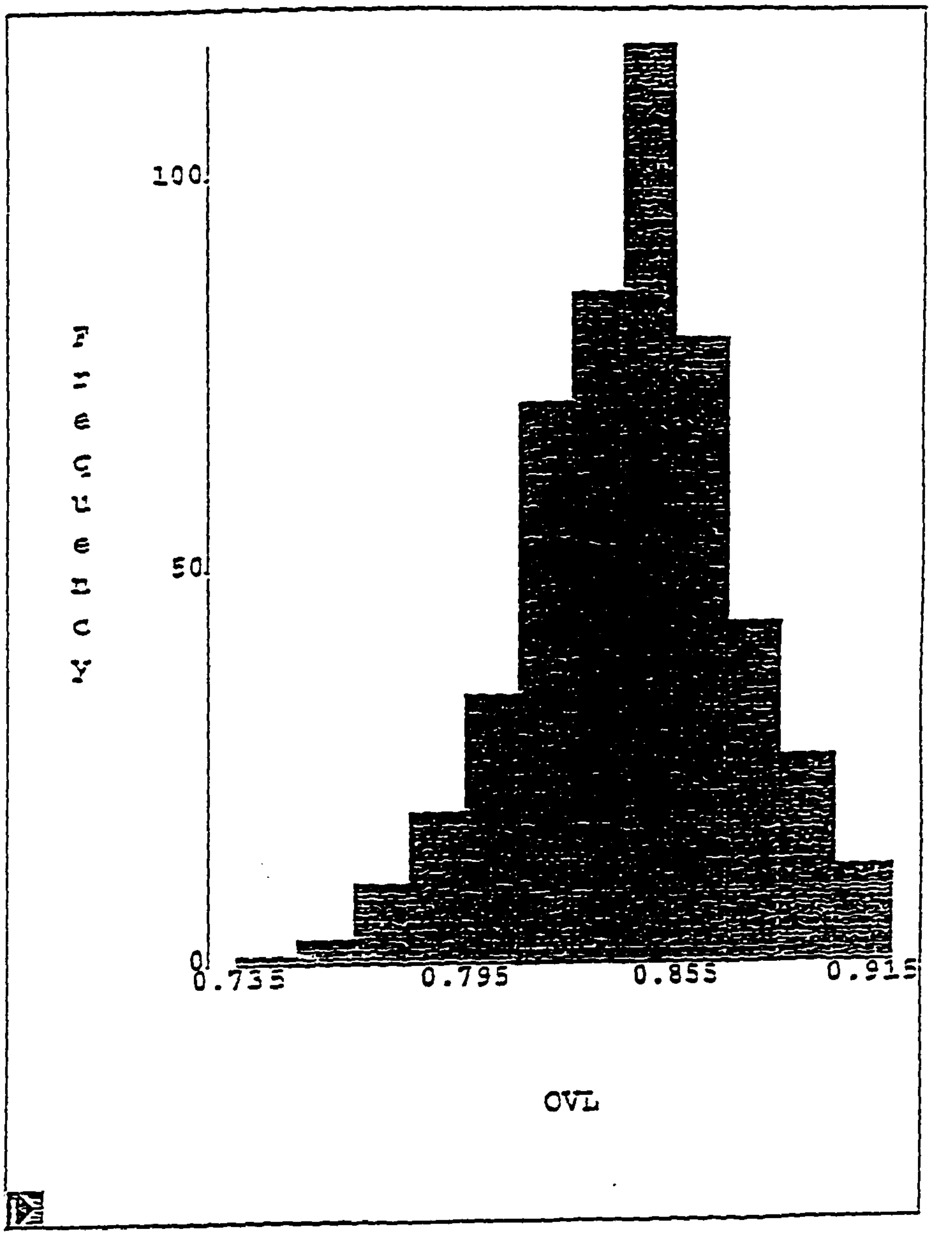

Figure 29. The histogram of the B = 500 bootstrap estimators for the Irish Education Transition data.

Using this data those patients experiencing chest pain for more than 6 hours before treatment were classified by gender. A Shapiro-Wilks test for normality of this data for the 131 males yielded a *p* value of 0.0001. The Shapiro-Wilks test for normality for the 69 females yielded a *p* value of 0.4358. Since the distribution for the minutes from onset of ischemic chest pain to ECG for males cannot be considered to follow a normal distribution, a nonparametric test was conducted to test differences in the median time (in minutes) from onset of ischemic chest pain to ECG between males and females. The median time for males is 508. While the median time for females is 565. A Median Two Sample Test using a normal approximation yields a z-statistic of 1.93862 which is moderately significant at the 0.0525 level. Thus it appears reasonable to assume that the median time (in minutes) from onset of ischemic chest pain to ECG differed significantly for male and females.

The bigger question might be to ask whether the distribution of the minutes from onset of ischemic chest pain to ECG differs for males and females. To make this comparison, we caculate the value of the overlapping coefficient. The $O\hat{V}L$ for the two distributions was 0.79234, which suggests that the distributions may not be as distinct as suggest by the simple comparison of medians (See Figures 30-33). The distribution of the time from onset of ischemic chest pain to ECG for males tended to be left skewed. The time from onset of ischemic chest pain to ECG for females tended to be concentrated near the median time. Thus the difference between the two distributions is more prominent the left tail of the distribution, while elsewhere the two distribitions tended to overlap.

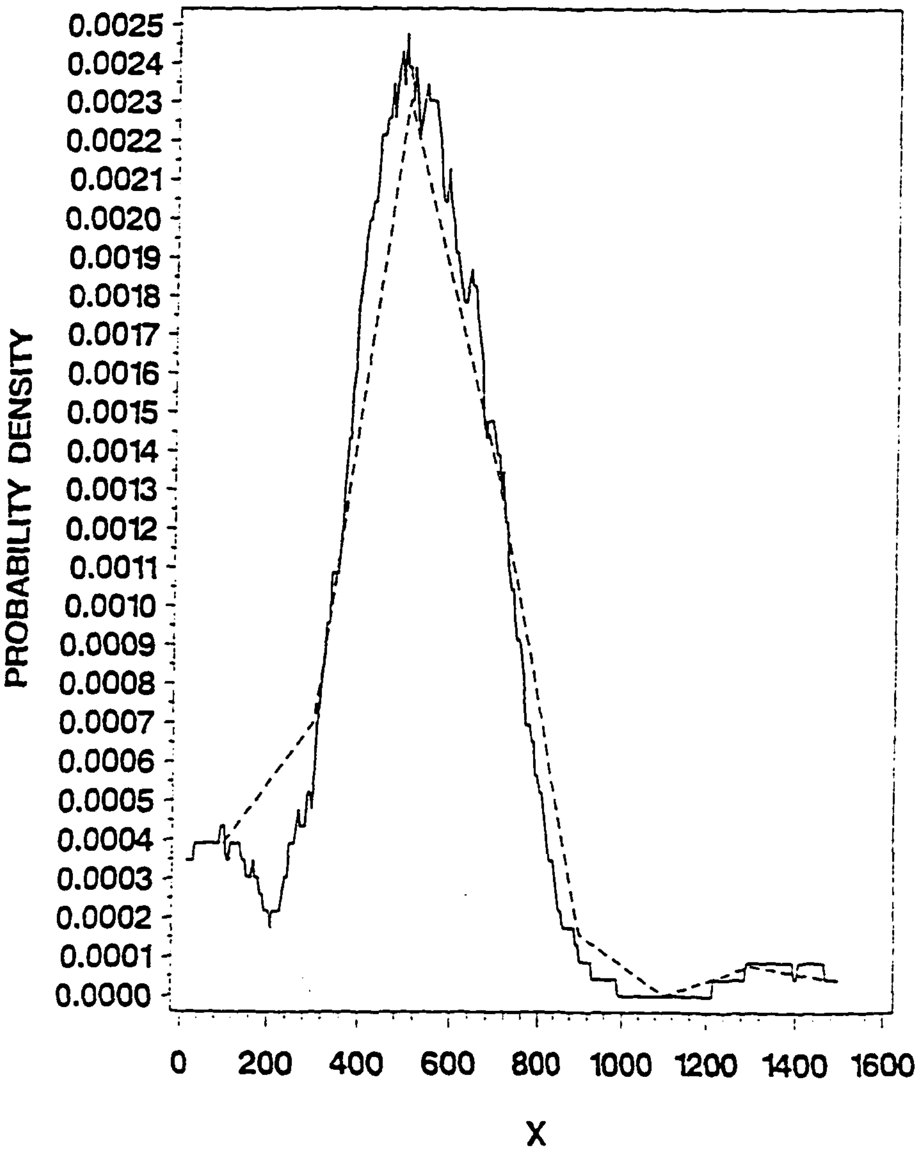

Figure 30. The kernel density estimator for the minutes from onset of ischemic chest pain to ECG for male patients. The dotted line denotes the midpoints of the histogram of the data. The solid line denotes the kernel density estimator for the data.

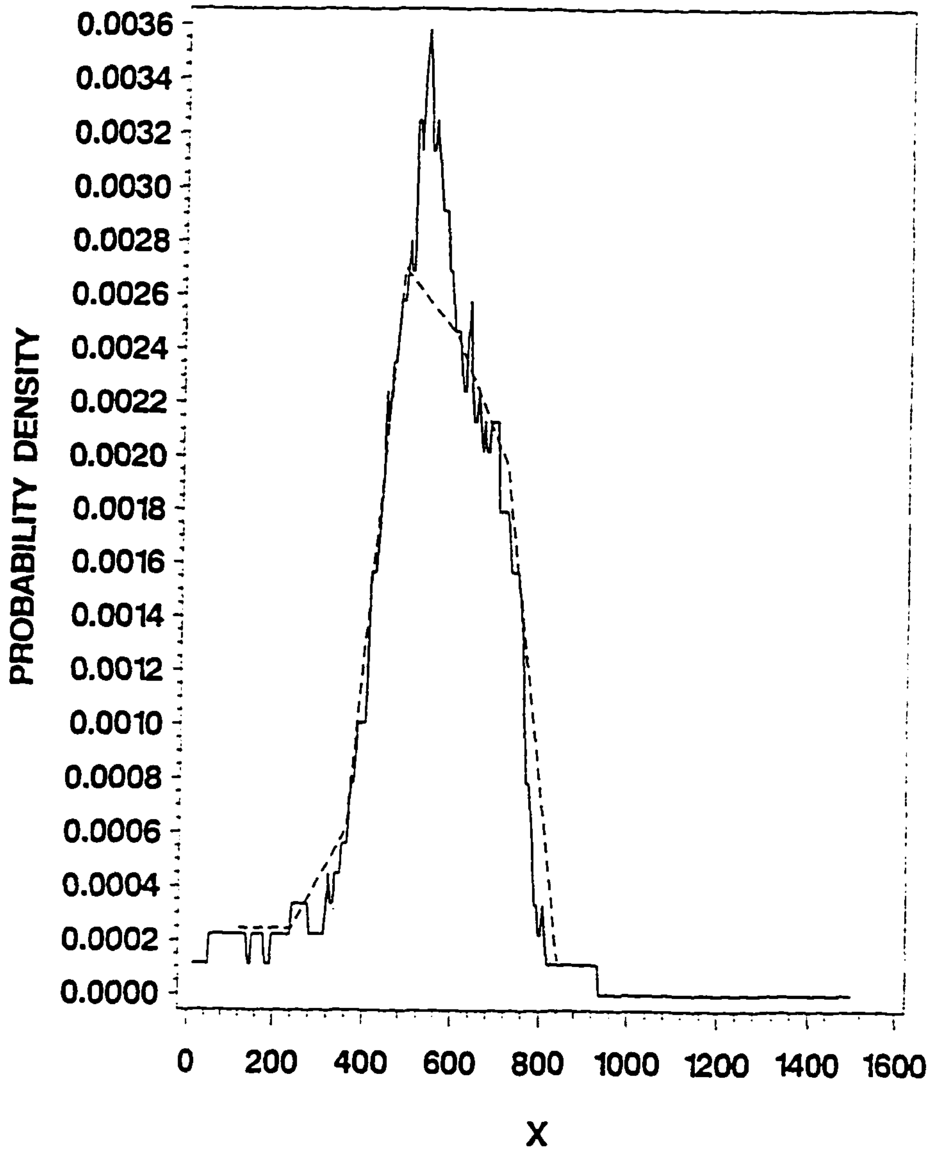

Figure 31. The kernel density estimator for the minutes from onset of ischemic chest pain for female patients. The dotted line denotes the midpoints of the histogram of the data. The solid line denotes the kernel density estimator for the data.

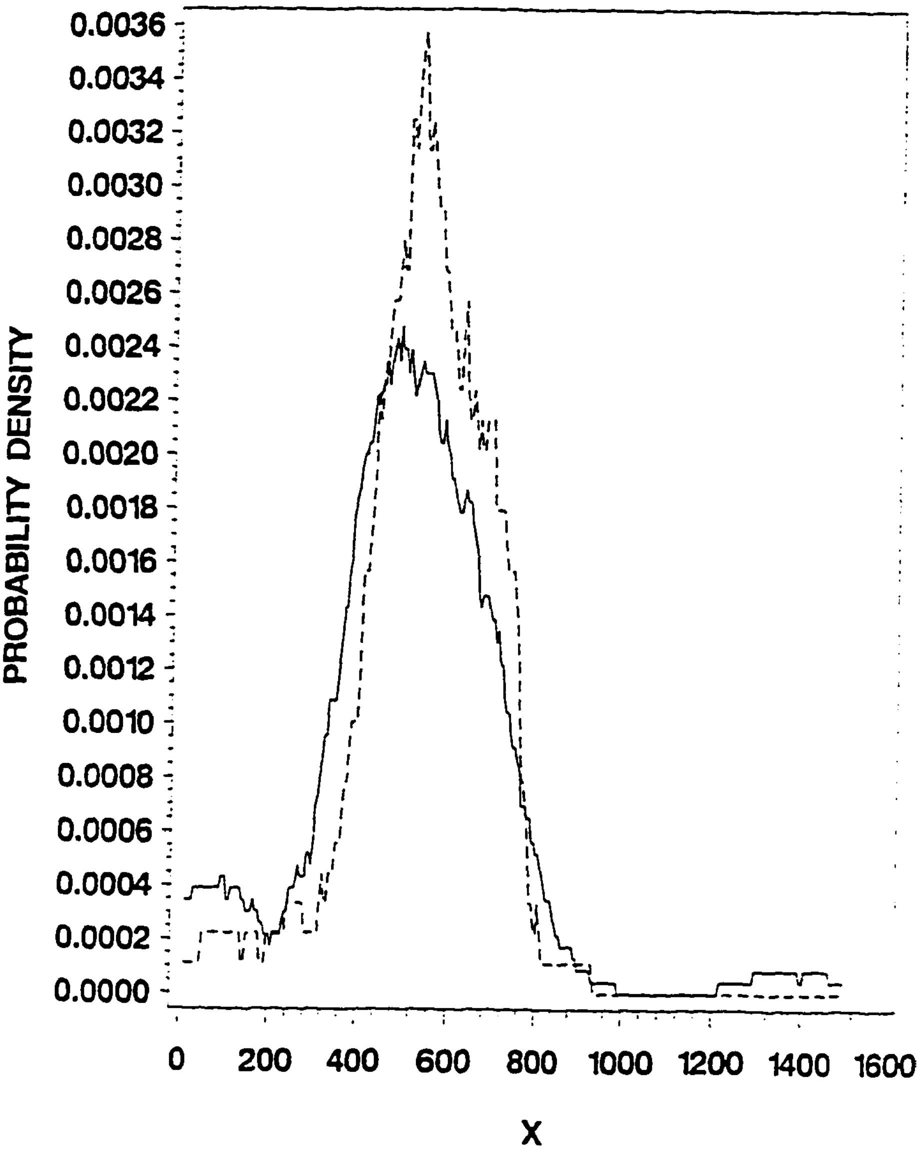

Figure 32. The kernel estimator of the overlap for the acute myocardial infarction data. The solid line denotes the kernel density estimator for the minutes from onset of ischemic chest pain for male patients. The dotted line denotes the kernel density estimator for the minutes from onset of ischemic chest pain for female patients.

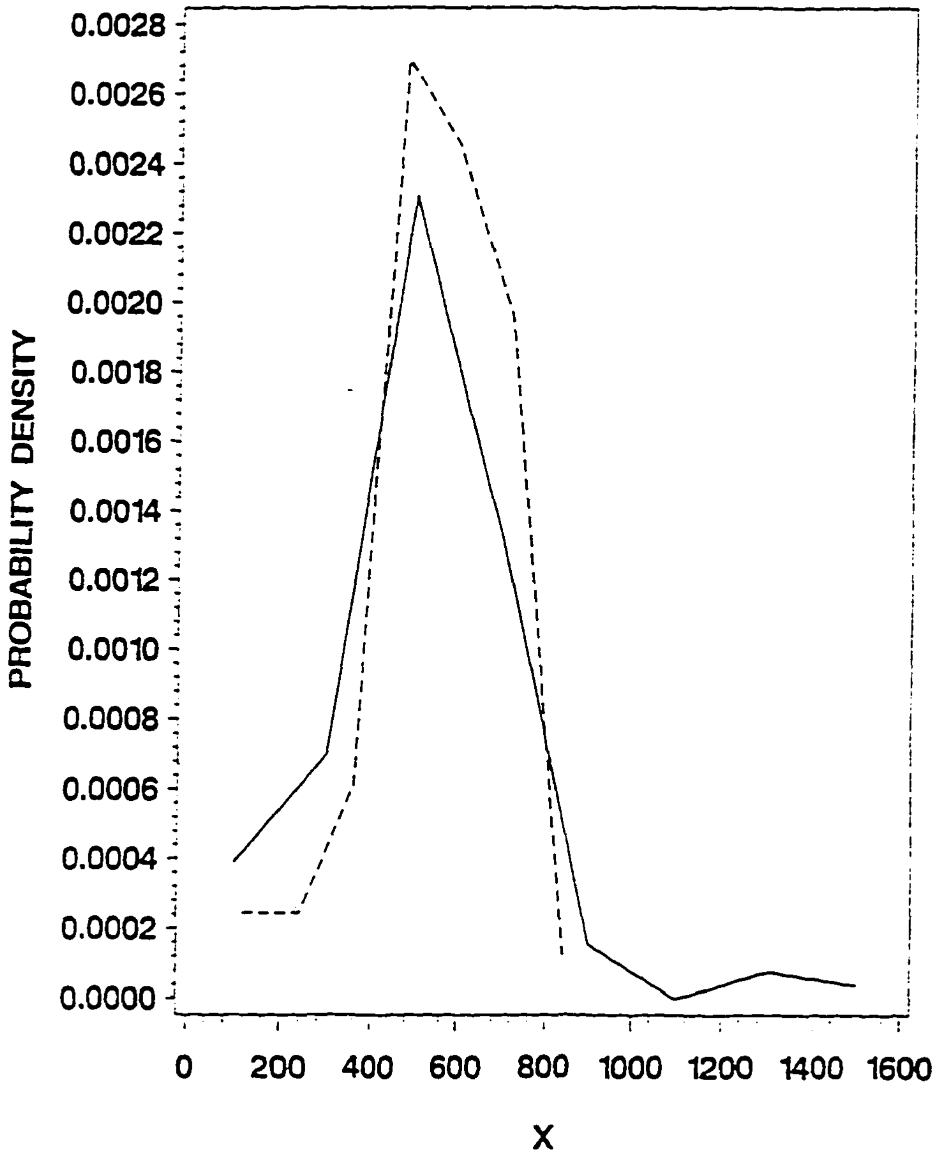

Figure 33. The overlap for the acute myocardial infarction data. The solid line denotes the midpoints for the histogram for the minutes from onset of ischemic chest pain for male patients. The dotted line denotes the midpoints for the histogram for the minutes from onset of ischemic chest pain for female patients.

Bootstrap estimates of the variance of the estimator of the OVL were obtained using Equation 34. Results for three different values of B are as follows:

$$\text{For B} = 100, \quad \tilde{Var}_B(O\tilde{V}L) = 0.0018893;$$

$$\text{For B} = 200, \quad \tilde{Var}_B(O\tilde{V}L) = 0.0015787;$$

$$\text{For B} = 500, \quad \tilde{Var}_B(O\tilde{V}L) = 0.0016212.$$

It must be noted that the variance increased slightly from B=200 to B=500. Using the results obtained when B = 500, the percentile method for constructing a bootstrap confidence interval for OVL described previously was used to compute a confidence interval for OVL using the 196 values. A 90% confidence interval for the true overlap between the time (in minutes) from onset of ischemic chest pain to ECG for females versus males who experienced chest pain for more than 6 hours before treatment, using the bootstrap distribution constructed from the 500 $O\tilde{V}L^{*}$, is given by

$$[F^{*}{}^{-1}_{500}(0.05), F^{*}{}^{-1}_{500}(0.95)] = (0.704385\,, 0.846907)\,.$$

Figure 34 is a histogram of the 500 bootstrap estimates. A Shapiro-Wilks test for normality was performed to test whether the empirical distribution of the bootstrap estimates can be considered normal. The p-value for the test was 0.1258. A summary of the results from the three examples is given in Appendix J.

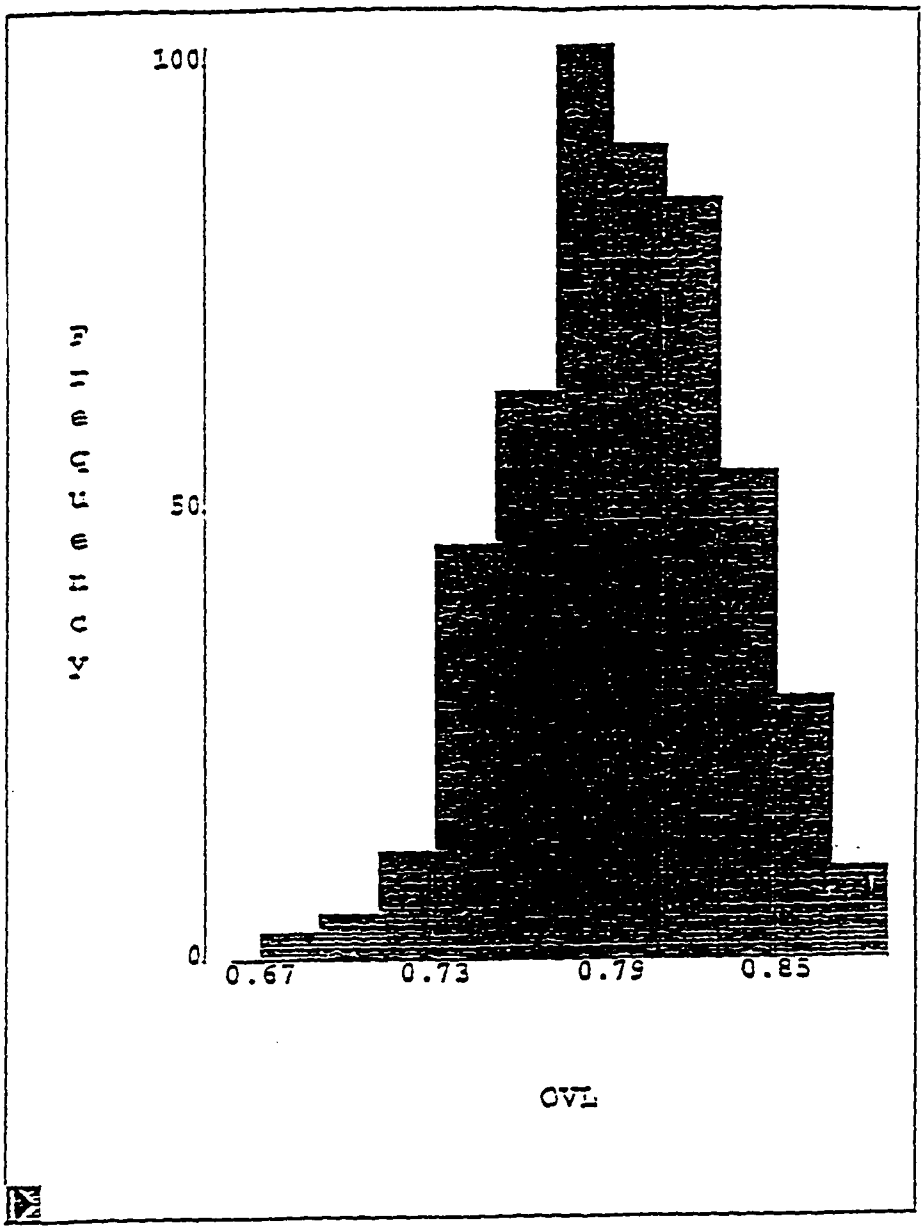

Figure 34. The histogram of the B = 500 bootstrap estimators for the acute myocardial infarction data.

V. CONCLUSION

The analysis of the behavior of the kernel estimator of the overlapping coefficient using the normal reference rule for the bandwidth for the naive kernel shows that the estimator of the OVL is a consistent estimator for the true overlap between two distributions. Also, the bootstrap variance estimator of the overlapping coefficient performs well as an estimator of variance.

The primary advantage of the kernel estimator of the overlapping coefficient is its distribution-free approach. In comparisons of the kernel estimator to the maximum likelihood estimator of the overlapping coefficient in the Normal distributional setting, it is seen that the kernel estimator should perform quite adequately in situations of more immediate interest, where the maximum likelihood estimator of the OVL would be inappropriate.

First, in evaluating the kernel density estimator using the normal reference rule, we have shown that the estimator (although it is one of the more simple kernel estimators, thus providing a rough estimator of the density of interest), tends to work adequately for all of the distributional settings used in the study. As for the kernel estimator of the overlapping coefficient, the estimator proves to be a biased estimator of the OVL where this bias is related to the value of the OVL and the sample size. With the exception of identical distributions, the mean bias was minimal. When sampling from two normal distributions the mean absolute relative bias of the kernel estimator of the overlapping coefficient was 1.5%

for $n = 100$ and 1.9% for $n = 500$; for two Weibull distributions the mean absolute relative bias was approximately 1.8% for $n = 100$ and 500; and for two gamma distributions the mean absolute relative bias was approximately 2.1% for $n = 100$ and 500. The estimator performed well when sampling from two beta distributions and mixtures of various distributions.

The bootstrap variance estimator also performed well with the exception of sampling from two identical distributions. For the case of equal distributions, the bootstrap estimator of variance greatly overestimated the variance with this bias being approximately 50% for each distributional-design-point-sample-size combination. For the case of distinct distributions, when sampling from two normal distributions, the mean absolute relative bias of the bootstrap estimator of the variance was 4.6% for $n = 100$ and 3.6% for $n = 500$; for two Weibull distributions the mean absolute relative bias of the bootstrap variance was 6.5% for $n = 100$ and 6.9% for $n = 500$; and for two gamma distributions the mean absolute relative bias was 7.3% for $n = 100$ and 4.5% for $n = 500$. The bootstrap estimator performed well when sampling from two beta distributions and mixtures of distributions.

As the examples indicate the kernel estimator of the OVL along with the bootstrap estimator of variance are efficient when dealing with real problems of data analysis where the distributional setting of the data is unknown. Also, it must be noted that the kernel density estimator using the normal reference rule performed reasonably well in estimating the densities of interest.

The properties of the kernel estimator of the OVL observed in the Monte Carlo experiment provide realistic guidance to the actual use of the estimator. In particular, the bias of the kernel estimator of the OVL and the problem of estimating its variance accurately circumscribe the use of the kernel estimator as an inferential statistic. Thus, the kernel estimator can best be used as a check of the meaningfulness of the differences in parameters that are detected using various non-parametric methods. Thus, the OVL offers a technique of exploring the meaningfulness of an apparent statistical difference between two distributions.

The disadvantage of using the OVL as a measure of association noted in Gastwirth (1975) is that the magnitude of the OVL does not indicate where the common probability mass is located. However, Inman and Bradley (1989) observed the OVL has some advantages compared to other measures of association. It offers a common approach for the measurement of agreement between two distributions in any distributional setting. Thus the OVL is less restrictive than other procedures keyed directly to distributional assumptions that may or may not prove warranted in data analysis. Also the OVL is based on a simple, easily comprehended concept of the association between two probability distributions. The OVL has an alternative interpretation based on the classification of individuals into two populations. Given the two distributions of the populations of interest, the OVL can be said to represent the sum of the conditional probabilities of misclassifying an individual into the two populations. The classification rule is the assignment of an individual at any level of the characteristic of concern to the population. In other words, the OVL is an indicator of the difference between individuals in two populations or the two distributions in general. Whether or not the OVL is useful in any situation depends of the meaning the OVL has in the context of a specific problem.

Further work in this area includes the development of conditional and/or unconditional tests of the overlapping coefficient using the maximum likelihood estimator of the reparameterized OVL. In addition, the development of tests for the OVL using the nonparametric estimator of the overlapping coefficient provides an area for further research.

REFERENCES

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society, Ser. B, 26, 211-252.

Bradley, E. L. & Piantadosi, S. (1982). The overlapping coefficient as a measure of agreement between distributions (Technical Report). Birmingham, AL: University of Alabama at Birmingham, Department of Biostatistics and Biomathematics.

Clemons, T. E. (1996). The overlapping coefficient for two normal probability functions with unequal variances. Unpublished master's thesis, University of Alabama at Birmingham.

Cortese, C. F., Falk, R. F. & Cohen, J. K. (1976). Further consideration of the methodological analysis of segregation indices. American Sociological Review, 41, 630-637.

de Boor, C. (1978). A practical guide to splines: Vol. 27, Applied Mathematical Sciences. New York: Spring-Verlag.

Duncan, O. D. & Duncan, B. (1955). A methodological analysis of segregation indices. American Sociological Review, 20, 210-217.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7, 1-26.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap, and other methods. Biometrika, 68, 589-599.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Efron, B. (1990). More efficient bootstrap computations. Journal of the American Statistical Association, 85, 79-89.

Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician, 37, 36-48.

Efron, B. & Tibshirani, R. J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 1, 54-77.

Efron, B. & Tibshirani, R. J. (1993). An introduction to the bootstrap. New York, NY: Chapman & Hall.

Fix, E. & Hodges, J. L. (1951). Nonparametric discrimination: Consistency properties (Report Number 4). USAF School of Aviation Medicine, Randolph Field, Texas.

Gastwirth, J. L. (1975). Measures of similarity, dissimilarity, and distance. Encyclopedia of Statistical Sciences, 5, 402-403.

Goodman, L. A. & Kruskall, W. H. (1979). Measures of association for cross classification. New York: Springer-Verlag.

Greaney, V. & Kelleghan, T. (1984). Equality of opportunity in Irish schools. Dublin: Educational Company.

Hamming, R. W. (1971). Introduction to applied numerical analysis. New York: McGraw-Hill.

Hogg, R. V. & Craig, A. T. (1970). Introduction to mathematical statistics (3rd ed.). London: Macmillan.

IMSL. (1991) IMSL library user's manual: Fortran subroutines for statistical analysis, ver.2. Houston: IMSL.

Inman, H. F. (1981). Migration in the cotton South: the geographic mobility of Alabama farmers, 1850-1860. Unpublished master's thesis. University of Alabama at Birmingham.

Inman, H. F. (1984). Behavior and properties of the overlapping coefficient as a measure of agreement between distributions. Unpublished doctoral dissertation, University of Alabama at Birmingham.

Inman, H. F. & Bradley, E. L. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Communications in Statistics--Theory and Methods, 18, 3851-3872.

Inman, H. F. & Bradley, E. L. (1991). Approximations to the mean and variance of the index of dissimilarity in 2xC tables under a random allocation model. Sociological Methods and Research, 20, 242-255.

Inman, H. F. & Bradley, E. L. (1994). Hypothesis tests and confidence interval estimates for the overlap of two normal distributions with equal variances. Environmetrics, 5, 167-189.

Johnson, N. L. & Kotz, S. (1970). Continuous univariate distributions, Vol. 2. New York: John Wiley.

Kendall, M. & Stuart, A. (1977). The advanced theory of statistics: Vol. 1 (4th ed.). New York: Macmillan.

Kendall, M. & Stuart, A. (1979). The advanced theory of statistics: Vol. 2 (4th ed.). New York: Macmillan.

Lindgren, B. W. (1968). Statistical Theory (2nd ed). London: Macmillan.

Marx, W. (1976). Die messung der assoziativen bedeutungsahnlichkeit. Zeitschrift fur Experimentelle und Angewandte Psychologie, 23, 62-76.

Marx, W. (1976). Die statistische sicherung des uberlappungs koeffizienten. Zeitschrift fur Experimentelle und Angewandte Psychologie, 23, 267-70.

Miller, A. R. (1982). Fortran programs for scientists and engineers. Berkeley: Sybex.

Mishra, S. N. & Mulekar, M. S. (1992, March). Bias and inference of measures of niche overlap. Paper presented at the Spring meeting of the International Biometrics Society, Cincinnati, OH.

Mishra, S. N., Shah, A. K. & Lefante, J. J. (1986). Overlapping coefficient: The generalized t approach. Communications in Statistics--Theory and Methods, 15, 123-128.

Montgomery, D. C. (1991). Design and analysis of experiments (3rd ed). New York: John Wiley & Sons.

Mulekar, M. S. & Mishra, S. N. (1992, March). Overlap coefficients of two normal densities: Equal means case. Paper presented at the Spring meeting of the International Biometrics Society, Cincinnati, OH.

Parzen, E. (1962). On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33, 1065-1076.

Pearson, K. (1895). Contributions to the mathematical theory of evolution II: Skew variations in homogeneous material. Philosophical Transactions of the Royal Society of London, Ser. A, 186, 343-414.

Rogers, W. J., Dean, L. S., Moore, P. B., Wool, K. J., Burgard, S. L. & Bradley, E. L. (1993). Comparison of primary angioplasty versus thrombolytic therapy for acute myocardial infarction. Birmingham, AL: Alabama Registry of Myocardian Ischemia Investigators.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics, 27, 832-837.

Scott, D. W. (1992). Multivariate Density Estimation--Theory, Practice and Visualization. New York: John Wiley.

Shapiro, S. S. & Wilks, M. B. (1965). An analysis of variance test for normality. Biometrika, 52, 591-611.

Silverman, B.W. (1986). Density estimation for statistics and data analysis. New York: Chapman & Hall.

Sneath, P. H. A. (1977). A method for testing the distinctness of clusters: A test of the disjunction of two clusters in euclidean space as measured by their overlap. Mathematical Geology, 9, 123-143.

Sneath, P. H. A. (1979). The sampling distribution of the W statistic of disjunction for the arbitrary division of a random rectangular distribution. Mathematical Geology, 11, 423-442.

Tukey, J. W. (1957). On the comparative anatomy of transformations. Annals of Mathematical Statistics, 28, 602-632.

Waterman, M. S. & Whiteman, D. E. (1978). Estimation of probability densities by empirical density functions. International Journal for Mathematical Education in Science and Technology, 9, 127-137.

Wegman, E. J. (1972). Nonparameteric probability density estimation: I. A summary of available methods. Technometrics, 14, 533-546.

Wegman, E. J. (1982). Density estimation. In Encyclopedia of statistical sciences (Vol. 2, pp. 309-15). New York: John Wiley.

Weitzman, M. S. (1970). Measure of overlap of income distribution of white and negro families in the United States (Technical Paper No. 22). Washington: U. S. Department of Commerce, Bureau of the Census.

APPENDIX A

FORTRAN SUBROUTINES USED IN THE MONTE CARLO SIMULATION INVESTAGATION OF THE KERNEL ESTIMATOR OF THE OVL

Subroutine MEANST

The subroutine MEANST (Miller, 1982) computes the mean and variance from a simple random sample.

```
SUBROUTINE MEANST(X,NX,U,V)
    DOUBLE PRECISION X(1)
    REAL U,V,SUM,SUMSQ
    INTEGER NX,I

    SUM=0.0D0
    SUMSQ=0.0D0
    DO 10 I=1,NX
    SUM=SUM+X(I)
    SUMSQ=SUMSQ+X(I)*X(I)
 10 CONTINUE
    U=SUM/NX
    V=((SUMSQ-SUM*SUM/NX)/(NX-1))
    RETURN
    END
```

Subroutine OVCOEF

The subroutine OVCOEF calculates the overlapping coefficient for the distributions of two sets of random variables by estimating the densities using the naive/Rosenblatt kernel estimator:

Called subroutines: VSRTD (IMSL), JUMPS, and INTERV (de Boor, 1978)

```
c*******************************************************************
 c*                                                                *
 c*    This routine is to calculate the overlapping coefficient for *
 c* the distributions of two sets of random variables by estimating *
 c* the densities and hence the OVC by means of the naive kernel    *
 c* estimator.                                                      *
 c*    The arguments of the routine have the following meanings:   *
 c*                                                                *
 c*    s1()......double precision array of size at least ns1 used to *
 c*            pass the data from sample 1 to the routine. On      *
 c*            exit from the routine the array is sorted in        *
 c*            ascending order.                                    *
 c*                                                                *
 c*    ns1.......integer variable used to pass the number of elements *
 c*            in s1() to the routine.                             *
 c*                                                                *
 c*    s2()......double precision array of size at least ns2 used to *
 c*            pass the data from sample 2 to the routine. On      *
 c*            exit from the routine the array is sorted in        *
 c*            ascending order.                                    *
 c*                                                                *
 c*    ns2.......integer variable used to pass the number of elements *
 c*            in s2() to the routine.                             *
 c*                                                                *
 c*    work()....double precision array which must have diemnsion  *
 c*            larger than 2(ns1+ns2). It is used internally in    *
 c*            the calculation of OVC.                             *
 c*                                                                *
 c*    h1,h2.....double precision variables which respectively are *
 c*            the step sizes for the kernel estimators for samples *
 c*            1 and 2.
    *
 c*                                                                *
 c*    ovc.......double precision variable used to return the value *
 c*            of the overlapping coefficient                      *
 c*                                                                *
 c* Latest Revision:  November 1996                                 *
 c*                                                                *
 c* Routines called:  VSRTD, JUMPS , INTERV                         *
 c*                                                                *
c*******************************************************************
 c*                                                                *
 c*         C A V I A T   R E C E P T O R
    *
 c*                                                                *
```

```
c*****************************************************************
      subroutine ovcoef(s1,ns1,s2,ns2,work,h1,h2,ovc)
      double precision s1(1),s2(1),work(1),h1,h2,ovc
      integer ns1,ns2
c
c           local variables
c
      double precision zero,half,one,x,y,intlen,f1(0:1),f2(0:1)
      double precision df1,df2
      integer npts1,npts2,left,mflag,i,j
      parameter(zero=0.0d0,half=0.5d0,one=1.0d0)
c
c
      open(10,file='scratch',status='unknown')
      call vsrtd(s1,ns1)
      call vsrtd(s2,ns2)
      call jumps(s1,ns1,h1,work(1),npts1)
        do i=1,npts1
        write(10,900) work(i)
        end do
      call jumps(s2,ns2,h2,work(npts1+1),npts2)
        do i=1,npts2
        write(10,900) work(npts1+i)
        end do
      npts1=npts1+npts2
      call vsrtd(work,npts1)
        do i=1,npts1
        write(10,900) work(i)
 900    format(1h ,1pd15.7)
        end do
        close(10,status='keep')
c
c          change points for each ECDF now found and merged in the
c          work storage array work(). Next calculate the OVC by
c          summing the area under the smaller curve over each
c          subinterval.
c
      ovc=zero
        do i=1,npts1-1
          intlen=work(i+1)-work(i)
          x=work(i)+half*intlen
          do j=0,1
            y=x+dfloat(2*j-1)*half*h1
            call interv(s1,ns1,y,left,mflag)
            select case(mflag)
```

```
              case(-1)
                f1(j)=zero
              case(0)
                if(y.eq.s1(ns1)) then
                  f1(j)=one
                else
                  f1(j)=dfloat(left)/dfloat(ns1)
                endif
              case(1)
                f1(j)=one
            end select
          end do
          df1=(f1(1)-f1(0))/h1
c
          do j=0,1
            y=x+dfloat(2*j-1)*half*h2
            call interv(s2,ns2,y,left,mflag)
            select case(mflag)
              case(-1)
                f2(j)=zero
              case(0)
                if(y.eq.s2(ns2)) then
                  f2(j)=one
                else
                  f2(j)=dfloat(left)/dfloat(ns2)
                endif
              case(1)
                f2(j)=one
            end select
          end do
          df2=(f2(1)-f2(0))/h2
c
          ovc=ovc+dmin1(df1,df2)*intlen
        end do
c
      return
      end
```

Subroutine VSRTD

The subroutine VSRTD (IMSL) sorts a given array by the algebraic values.

```
C  IMSL ROUTINE NAME   - VSRTD
   C
   C-----------------------------------------------------------
   C
   C  COMPUTER          - IBM/SINGLE
   C
   C  LATEST REVISION    - JANUARY 1, 1978
   C
   C  PURPOSE           - SORTING OF ARRAYS BY ALGEBRAIC VALUE
   C
   C  USAGE             - CALL VSRTA (A,LA)
   C
   C  ARGUMENTS   A     - ON INPUT, A CONTAINS THE ARRAY TO BE
       SORTED.

   C                  ON OUTPUT, A CONTAINS THE SORTED ARRAY.
   C              LA    - INPUT VARIABLE CONTAINING THE NUMBER OF
   C                     ELEMENTS IN THE ARRAY TO BE SORTED.
   C
   C  PRECISION/HARDWARE  - DOUBLE/ALL
   C
   C  REQD. IMSL ROUTINES - NONE REQUIRED
   C
   C  NOTATION           - INFORMATION ON SPECIAL NOTATION AND
   C                     CONVENTIONS IS AVAILABLE IN THE MANUAL
   C                     INTRODUCTION OR THROUGH IMSL ROUTINE UHELP
   C
   C  COPYRIGHT          - 1978 BY IMSL, INC. ALL RIGHTS RESERVED.
   C
   C  WARRANTY           - IMSL WARRANTS ONLY THAT IMSL TESTING HAS
   C  BEEN
   C                     APPLIED TO THIS CODE. NO OTHER WARRANTY,
   C                     EXPRESSED OR IMPLIED, IS APPLICABLE.
   C
   C-----------------------------------------------------------
   C
      SUBROUTINE VSRTD  (A,LA)
   C                        SPECIFICATIONS FOR ARGUMENTS
      INTEGER          LA
      DOUBLE PRECISION   A(LA)
   C                        SPECIFICATIONS FOR LOCAL VARIABLES
      INTEGER          IU(21),IL(21),I,M,J,K,IJ,L
      DOUBLE PRECISION   T,TT,R
   C                        FIRST EXECUTABLE STATEMENT
      M=1
      I=1
      J=LA
      R=.375D0
      IF (LA.LE.0) RETURN
   10 IF (I .EQ. J) GO TO 55
   15 IF (R .GT. .5898437D0) GO TO 20
      R=R+3.90625D-2
      GO TO 25
   20 R=R-.21875D0
   25 K=I
C                        SELECT A CENTRAL ELEMENT OF THE
      IJ=I+(J-I)*R
      T=A(IJ)
C                        IF FIRST ELEMENT OF ARRAY IS GREATER
C                        THAN T, INTERCHANGE WITH T
      IF (A(I) .LE. T) GO TO 30
      A(IJ)=A(I)
      A(I)=T
      T=A(IJ)
   30 L=J
C                        IF LAST ELEMENT OF ARRAY IS LESS THAN
C                        T, INTERCHANGE WITH T
      IF (A(J) .GE. T) GO TO 40
      A(IJ)=A(J)
      A(J)=T
      T=A(IJ)
C                        IF FIRST ELEMENT OF ARRAY IS GREATER
C                        THAN T, INTERCHANGE WITH T
      IF (A(I) .LE. T) GO TO 40
      A(IJ)=A(I)
      A(I)=T
      T=A(IJ)
      GO TO 40
   35 IF(A(L).EQ.A(K)) GO TO 40
      TT=A(L)
      A(L)=A(K)
      A(K)=TT
C                        FIND AN ELEMENT IN THE SECOND HALF OF
C                        THE ARRAY WHICH IS SMALLER THAN T
   40 L=L-1
      IF (A(L) .GT. T) GO TO 40
C                        FIND AN ELEMENT IN THE FIRST HALF OF
C                        THE ARRAY WHICH IS GREATER THAN T
   45 K=K+1
      IF (A(K) .LT. T) GO TO 45
```

```
C                     INTERCHANGE THESE ELEMENTS
      IF (K .LE. L) GO TO 35
C                     SAVE UPPER AND LOWER SUBSCRIPTS OF
C                     THE ARRAY YET TO BE SORTED
      IF (L-I .LE. J-K) GO TO 50
      IL(M)=I
      IU(M)=L
      I=K
      M=M+1
      GO TO 60
   50 IL(M)=K
      IU(M)=J
      J=L
      M=M+1
      GO TO 60
C                     BEGIN AGAIN ON ANOTHER PORTION OF
C                     THE UNSORTED ARRAY
   55 M=M-1
      IF (M .EQ. 0) RETURN
      I=IL(M)
      J=IU(M)
   60 IF (J-I .GE. 11) GO TO 25
      IF (I .EQ. 1) GO TO 10
      I=I-1
   65 I=I+1
      IF (I .EQ. J) GO TO 55
      T=A(I+1)
      IF (A(I) .LE. T) GO TO 65
      K=I
   70 A(K+1)=A(K)
      K=K-1
      IF (T .LT. A(K)) GO TO 70
      A(K+1)=T
      GO TO 65
      END
```

Subroutine JUMPS

The subroutine JUMPS locates the points at which the naive/Rosenblatt kernel density estimator has jumps.

Called subroutines: INTERV

```
c******************************************************************
  c*                                                              *
  c*    The purpose of this routine is to locate the points at which
  c*  the naieve kernel density estimator has jumps. The routine assumes
  c*  that the vector of observations passed to the routine are sorted
  c*  from smallest to largest. The arguments of the routine have the
  c*  following meanings:
  c*
  c*     x().......double precision vector of observations upon which
  c*               the empirical distribution function is based. This
  c*               vector is on length nx and is assumed to be sorted
  c*               in ascending order.
  c*
  c*     nx........integer variable used to tell the routine how many
  c*               elements there are in the vector x().
  c*
  c*     h.........double precision variable used to define the step
  c*               size used by the naieve kernel estimator. This

  c*               routine assumes the formula
  c*                 f(x)=[F(x+h/2) - F(x-h/2)]/h
  c*
  c*     wk()......double precision vector of length at least 2*nx
  c*               in which the routine will return the jump points
  c*               of the kernel estimator.
  c*
  c*     npts......integer variable in which the routine will return
  c*               return the number of jump points in the vector wk()
  c*
  c* Latest Revision:   November 1996
  c*
  c* Routines called:   interval.for
  c*
  c******************************************************************
  c*
  c*          C A V I A T   R E C E P T O R

  c*
  c******************************************************************
  c
      subroutine jumps(x,nx,h,wk,npts)
      double precision x(1),wk(1),h
      integer nx,npts
  c
  c                 local variables
```

```
c
      double precision lower,upper,halfh,dl,du
       integer mflagl,leftl,mflagu,leftu
c
c
       npts=0
       halfh=0.5d0*h
       upper=x(1)-halfh
       lower=upper-h
       call interv(x,nx,upper,leftu,mflagu)
       call interv(x,nx,lower,leftl,mflagl)
c
c              start main loop
c
      do while (lower.lt.x(nx))
        select case(mflagl)
          case(-1)
            dl=x(1)-lower
          case(0)
            dl=x(leftl+1)-lower
          case(1)
            stop 'TERMINAL ERROR...lower larger than x(nx)'
        end select
        select case(mflagu)
          case(-1)
            du=x(1)-upper
          case(0)
            if(upper.lt.x(nx)) then
              du=x(leftu+1)-upper
            else
              du=1.0d+200
            endif
          case(1)
            du=1.0d+200
        end select
        if(du .le. dl) then
          if(mflagu.eq.-1) then
            upper=x(1)
          else
            upper=x(leftu+1)
          endif
          npts=npts+1
          if(npts.gt.2*nx) stop 'Terminal error...wk() too large'
          wk(npts)=upper-halfh
          lower=lower+du
        else
          if(mflagl.eq.-1) then
            lower=x(1)
          else
            lower=x(leftl+1)
          endif
          npts=npts+1
          if(npts.gt.2*nx) stop 'Terminal error...wk() too large'
          wk(npts)=lower+halfh
          upper=upper+dl
        endif
        call interv(x,nx,upper,leftu,mflagu)
        call interv(x,nx,lower,leftl,mflagl)
      end do
c
c
      return
      end
```

Subroutine INTERV

The subroutine INTERV, from de Boor (1978), computes the interval between consecutive jump points.

```
subroutine interv ( xt, lxt, x, left, mflag )
 c  from  * a practical guide to splines *  by C. de Boor
 computes  left = max( i :  xt(i) .lt. xt(lxt) .and.  xt(i) .le. x ) .
 c
 c****** i n p u t ******
 c  xt.....a real sequence, of length  lxt , assumed to be nondecreasing
 c  lxt.....number of terms in the sequence  xt .
 c  x.....the point whose location with respect to the sequence  xt  is
 c       to be determined.
 c
 c****** o u t p u t ******
 c  left, mflag.....both integers, whose value is
 c
 c  1    -1    if             x .lt.  xt(1)
 c  i     0    if  xt(i)  .le. x .lt. xt(i+1)
 c  i     0    if  xt(i)  .lt. x .eq. xt(i+1) .eq. xt(lxt)
 c  i     1    if  xt(i)  .lt.        xt(i+1) .eq. xt(lxt) .lt. x
 c
 c       In particular,  mflag = 0  is the 'usual' case.  mflag .ne. 0
 c       indicates that  x  lies outside the CLOSED interval
 c       xt(1) .le. y .le. xt(lxt) . The asymmetric treatment of the
 c       intervals is due to the decision to make all pp functions cont-
 c       inuous from the right, but, by returning  mflag = 0  even if
 C        x = xt(lxt), there is the option of having the computed pp function
 c       continuous from the left at  xt(lxt) .
 c
 c****** m e t h o d ******
 c  The program is designed to be efficient in the common situation that
 c  it is called repeatedly, with  x  taken from an increasing or decrea-
 c  sing sequence. This will happen, e.g., when a pp function is to be
 c  graphed. The first guess for  left  is therefore taken to be the val-
 c  ue returned at the previous call and stored in the  l o c a l  varia-
 c  ble  ilo . A first check ascertains that  ilo .lt. lxt (this is nec-
 c  essary since the present call may have nothing to do with the previ-
 c  ous call). Then, if  xt(ilo) .le. x .lt. xt(ilo+1), we set  left =
 c  ilo  and are done after just three comparisons.
 c     Otherwise, we repeatedly double the difference  istep = ihi - ilo
 c  while also moving  ilo  and  ihi  in the direction of  x , until
 c                    xt(ilo) .le. x .lt. xt(ihi) ,
 c  after which we use bisection to get, in addition, ilo+1 = ihi .
 c  left = ilo  is then returned.
 c
     integer left,lxt,mflag,   ihi,ilo,istep,middle
     double precision x,xt(lxt)
     save ilo
```

```
      data ilo /1/
      ihi = ilo + 1
      if (ihi .lt. lxt)            go to 20
        if (x .ge. xt(lxt))          go to 110
        if (lxt .le. 1)            go to 90
        ilo = lxt - 1
        ihi = lxt
c
   20 if (x .ge. xt(ihi))            go to 40
      if (x .ge. xt(ilo))           go to 100
c
c              **** now x .lt. xt(ilo) . decrease  ilo  to capture  x .
      istep = 1
   31    ihi = ilo
         ilo = ihi - istep
         if (ilo .le. 1)           go to 35
         if (x .ge. xt(ilo))         go to 50
         istep = istep*2
                                 go to 31
   35 ilo = 1
      if (x .lt. xt(1))            go to 90
                                 go to 50
c              **** now x .ge. xt(ihi) . increase  ihi  to capture  x .
   40 istep = 1
   41    ilo = ihi
         ihi = ilo + istep
         if (ihi .ge. lxt)          go to 45
         if (x .lt. xt(ihi))         go to 50
         istep = istep*2
                                 go to 41
   45 if (x .ge. xt(lxt))            go to 110
      ihi = lxt
c
c           **** now xt(ilo) .le. x .lt. xt(ihi) . narrow the interval.
   50 middle = (ilo + ihi)/2
      if (middle .eq. ilo)          go to 100
c     note. it is assumed that middle = ilo in case ihi = ilo+1 .
      if (x .lt. xt(middle))          go to 53
        ilo = middle
                                 go to 50
   53    ihi = middle
                                 go to 50
c**** set output and return.
   90 mflag = -1
      left = 1
```

```
                              return
100 mflag = 0
   left = ilo
                              return
110 mflag = 1
   if(x .eq. xt(lxt)) mflag = 0
   left = lxt
111 if(left .eq. 1)                 return
   left = left - 1
   if(xt(left) .lt. xt(lxt))       return
                              go to 111
   end
```

APPENDIX B

DESIGN POINTS

Table B1

The Twelve Normal Distribution Design Points Used for the Simulation Study

mean 1	variance 1	mean 2	variance 2	OVL
0	1	0	1	1.00000
2	4	2	4	1.00000
0	1	3	5	0.31532
0	1	1	1	0.61708
0	1	0	3	0.74064
2	4	0	1	0.45339
1	1	0	3	0.63943
5	10	3	5	0.68421
2	4	3	5	0.80847
0	3	2	4	0.58875
1	1	3	5	0.45740
1	1	2	4	0.60993

Table B2

The Twelve Gamma Distribution Design Points Used for the Simulation Study

alpha 1	alpha 2	OVLl
1.5	2.5	0.69163957
1.5	2.0	0.83061704
1.5	4.0	0.40174569
1.5	3.5	0.48122526
2.5	2.0	0.85447962
2.5	4.0	0.65488521
2.5	3.5	0.75591698
2.0	4.0	0.52950408
2.0	3.5	0.62178956
4.0	3.5	0.89132310
1.5	1.5	1.00000000
4.0	4.0	1.00000000

Table B3

The Twelve Weibull Distribution Design Points Used for the Simulation Study

alpha 1	betal 1	alpha 1	beta 2	OVL
1.5	4.0	2.0	2.0	0.64209627
1.5	1.5	1.5	4.0	0.57323190
2.0	3.0	2.0	2.0	0.80280065
3.0	1.5	2.0	2.0	0.73724924
1.5	1.5	2.0	2.0	0.85696873
3.0	1.5	1.5	4.0	0.42141420
3.0	1.5	3.0	1.5	1.00000000
2.0	2.0	2.0	2.0	1.00000000
1.0	1.5	1.0	2.0	0.867784119
1.0	1.5	1.0	3.0	0.69097628
1.0	2.0	1.0	3.0	0.81489838
1.0	2.0	1.0	3.5	0.74739131

Table B4

The Four Beta Distribution Design Points Used for the Simulation Study

alpha 1	beta 1	alpha 2	beta 2	OVL
2	2	1	1	0.80755008
3	3	2	2	0.89266882
5	3	3	3	0.72247886
5	3	5	3	1.00000000

APPENDIX C

RESULTS OF THE MONTE CARLO SIMULATION STUDY: THE KERNEL ESTIMATOR OF THE OVL USING THE NORMAL REFERENCE RULE

Table C1

Results of the Monte Carlo Simulation Study: The Kernel Estimator of OVL Based on Independent Samples from Two Normal Distributions

	Predicted	Monte Carlo		Standard	Variance	Relative Bias	Relative Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
			$\mu_1 = 0$, $\sigma^2_1 = 1$, $\mu_2 = 0$ and $\sigma^2_2 = 1$, OVL = 1.000000				
100	0.0013639	0.89156	0.00092446	-3.56655	0.6778014	-10.84404	47.53584
500	0.0002742	0.951053	0.00017599	- 3.68967	0.6418855	-4.89475	55.79102
			$\mu_1 = 2$, $\sigma^2_1 = 4$, $\mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 1.000000				
100	0.00139068	0.892119	0.00090138	-3.59327	0.6481556	-10.78806	54.28394
500	0.00025195	0.950718	0.00016714	-3.81198	0.6633796	-4.92824	50.74324
			$\mu_1 = 2$, $\sigma^2_1 = 4$, $\mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.808473				
100	0.00266908	0.796343	0.00287713	-0.22615	1.0779476	-1.50039	- 7.23111
500	.000628652	0.811475	0.00064401	-0.11563	1.0721509	0.37132	- 6.72955
			$\mu_1 = 0$, $\sigma^2_1 = 1$, $\mu_2 = 0$ and $\sigma^2_2 = 3$, OVL = 0.740641				
100	0.00234662	0.724487	0.0025474	-3.20072	1.0855615	-2.18116	- 7.88177
500	0.00054237	0.736323	0.00057393	-0.01802	1.0581726	-0.58297	- 5.49746

(table continues)

						Relative	Relative
	Predicted	Monte Carlo		Standard	Variance	Bias	Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
$\mu_1 = 5, \sigma^2_1 = 10, \mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.683020							
100	0.00264340	0.674443	0.00289080	- 0.15952	1.0935910	- 1.25575	- 8.55813
500	.000581291	0.688872	0.00058938	0.24105	1.0139174	0.85679	- 1.37264
$\mu_1 = 1, \sigma^2_1 = 1, \mu_2 = 0$ and $\sigma^2_2 = 3$, OVL = 0.639430							
100	0.00247203	0.631868	0.00253316	- 3.20072	1.0247273	- 1.18255	- 2.41306
500	0.00537266	0.644058	0.00053212	0.20065	1.0682326	0.72388	- 6.38743
$\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 1$ and $\sigma^2_2 = 1$, OVL = 0.617075							
100	0.00316916	0.624577	0.00312376	0.13422	0.9678459	1.21568	3.32224
500	0.00061978	0.630828	0.00062559	0.54984	1.0093701	2.22868	- 0.92831
$\mu_1 = 1, \sigma^2_1 = 1, \mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 0.609934							
100	0.00235149	0.602939	0.00238396	- 0.14328	1.0138089	- 1.14697	- 1.36209
500	0.00050378	0.613520	0.00054184	0.15406	1.0755569	0.58794	- 7.02491

(table continues)

						Relative	Relative
	Predicted	Monte Carlo		Standard	Variance	Bias	Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
			$\mu_1 = 0, \sigma^2_1 = 3, \mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 0.588750				
100	0.00280185	0.599083	0.00295912	0.18996	1.0561286	1.75518	- 5.31456
500	0.00059996	0.603017	0.00057122	0.59696	0.9823646	2.42336	1.79520
			$\mu_1 = 1, \sigma^2_1 = 1, \mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.457402				
100	0.00214170	0.460238	0.00226334	0.05961	1.0567947	0.61980	- 5.37424
500	0.00044314	0.468227	0.00043951	0.51637	0.9920800	2.36671	- 0.79832
			$\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 0.453388				
100	0.00222756	0.459458	0.00214563	0.13103	0.9632191	1.33870	3.81854
500	0.00046184	0.466920	0.00044598	0.64075	0.9656596	2.98452	3.55616
			$\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.315318				
100	0.00177039	0.325574	0.00178051	0.24305	1.0057169	3.25248	- 0.56844
500	0.00368558	0.332589	0.00037529	- 0.89152	1.0182619	5.47727	- 1.79344

Table C2

Results of the Monte Carlo Simulation Study: The Kernel Estimator of OVL Based on Independent Samples from Two Weibull Distributions

		Monte Carlo				Relative	Relative
	Predicted			Standard	Variance	Bias	Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
			$\alpha_1 = 3.0$, $\beta_1 = 1.5$, $\alpha_2 = 3.0$ and $\beta_2 = 1.5$, OVL = 1.000000				
100	0.00134596	0.895581	0.00086694	- 3.546364	0.6441050	-10.44185	55.25419
500	0.00026180	0.953311	0.00017337	- 3.545980	0.6621966	- 4.66893	51.01256
			$\alpha_1 = 2.0$, $\beta_1 = 2.0$, $\alpha_2 = 2.0$ and $\beta_2 = 2.0$, OVL = 1.000000				
100	0.00135755	0.894509	0.00087026	- 3.575953	0.6410524	-10.54914	55.99348
500	0.00026572	0.952482	0.00017360	- 3.606483	0.6533218	- 4.75181	53.06393
			$\alpha_1 = 1.0$, $\beta_1 = 1.5$, $\alpha_2 = 1.0$ and $\beta_2 = 2.0$, OVL = 0.867841				
100	0.00180877	0.844289	0.00184481	- 0.548357	1.0227823	- 2.71393	- 2.22748
500	0.00049736	0.868469	0.00053016	- 0.027264	1.0659471	0.07234	- 6.18672
			$\alpha_1 = 1.5$, $\beta_1 = 1.5$, $\alpha_2 = 2.0$ and $\beta_2 = 2.0$, OVL = 0.856969				
100	0.00175138	0.856243	0.00182789	- 0.016976	1.0436842	- 0.08469	- 4.18558
500	0.00047459	0.888312	0.00050382	1.396392	1.0615973	3.65746	- 5.80233

(table continues)

						Relative	Relative
	Predicted	Monte Carlo		Standard	Variance	Bias	Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
$\alpha_1 = 1.0, \beta_1 = 2.0, \alpha_2 = 1.0$ and $\beta_2 = 3.0$, OVL = 0.814898							
100	0.00213969	0.796958	0.00234121	- 0.370774	1.0941807	- 2.20154	- 8.60742
500	0.00055286	0.817126	0.00057163	0.093170	1.0339482	2.73357	- 3.28335
$\alpha_1 = 2.0, \beta_1 = 3.0, \alpha_2 = 2.0$ and $\beta_2 = 2.0$, OVL = 0.802801							
100	0.00208043	0.797459	0.00229485	- 0.111515	1.1030646	- 0.66543	- 9.34348
500	0.00052013	0.815774	0.00059749	0.530744	1.1487463	1.61601	-12.94858
$\alpha_1 = 1.0, \beta_1 = 2.0, \alpha_2 = 1.0$ and $\beta_2 = 3.5$, OVL = 0.747391							
100	0.00234276	0.738898	0.00273571	- 0.162375	1.1677294	- 1.13634	-14.36373
500	0.00055134	0.751224	0.00056605	0.161107	1.0266937	0.51286	- 2.59997
$\alpha_1 = 3.0, \beta_1 = 1.5, \alpha_2 = 2.0$ and $\beta_2 = 2.0$, OVL = 0.737249							
100	0.00276862	0.725117	0.00313173	- 0.216789	1.1311516	- 1.64556	-11.59452
500	0.00066599	0.735910	0.00071848	- 0.049951	1.0789804	- 0.18161	- 7.30600

(table continues)

						Relative	Relative
	Predicted	Monte Carlo		Standard	Variance	Bias	Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
$\alpha_1 = 1.0, \beta_1 = 1.5, \alpha_2 = 1.0$ and $\beta_2 = 3.0$, OVL = 0.690976							
100	0.00229675	0.692287	0.00234908	0.270444	1.0199242	0.18970	- 2.22748
500	0.00050670	0.699696	0.00052742	0.379681	1.0408907	1.26192	- 3.92843
$\alpha_1 = 1.5, \beta_1 = 4.0, \alpha_2 = 2.0$ and $\beta_2 = 2.0$, OVL = 0.642096							
100	0.00273173	0.624900	0.00287333	- 0.320799	1.0518359	- 2.67809	- 4.92813
500	0.00051202	0.633298	0.00060685	- 0.357147	1.1852077	- 1.37021	-15.66601
$\alpha_1 = 1.5, \beta_1 = 1.5, \alpha_2 = 1.5$ and $\beta_2 = 4.0$, OVL = 0.573232							
100	0.00216456	0.581249	0.00234628	0.165510	1.0839562	1.39857	- 7.74535
500	0.00046681	0.587127	0.00045128	0.654070	0.9667304	2.42391	3.44146
$\alpha_1 = 3.0, \beta_1 = 1.5, \alpha_2 = 1.5$ and $\beta_2 = 4.0$, OVL = 0.421414							
100	0.00234763	0.399402	0.00235331	- 0.453758	1.0024209	- 5.22342	- 0.24151
500	0.00046199	0.404222	0.00050245	- 0.766997	1.0875798	- 4.07973	- 8.05273

Table C3

Results of the Monte Carlo Simulation Study: The Kernel Estimator of OVL Based on Independent Samples from Two Gamma Distributions

	Predicted	Monte Carlo		Standard	Variance	Relative Bias	Relative Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
			$\alpha_1 = 1.5, \alpha_2 = 1.5$, OVL = 1.000000				
100	0.00135007	0.894091	0.00087504	- 3.580311	0.6481425	-10.59094	54.28705
500	0.00027304	0.952504	0.00019297	- 3.419126	0.7067401	- 4.74963	41.49474
			$\alpha_1 = 4.0, \alpha_2 = 4.0$, OVL = 1.000000				
100	0.00132939	0.894121	0.00088684	- 3.555366	0.6671053	-10.58786	49.90137
500	0.00026669	0.951625	0.00017396	- 3.667674	0.6522953	- 4.83745	53.30480
			$\alpha_1 = 4.0, \alpha_2 = 3.5$, OVL = 0.891323				
100	0.00173601	0.857861	0.00195780	- 0.756264	1.1277583	- 3.75425	-11.32852
500	0.00054142	0.890024	0.00056521	- 0.546267	1.0439343	- 0.14570	- 4.20853
			$\alpha_1 = 2.5, \alpha_2 = 2.0$, OVL = 0.854480				
100	0.00202232	0.833808	0.00215044	- 0.445768	1.0633557	- 2.41919	- 5.95809
500	0.00060572	0.858150	0.00068788	0.139957	1.1356381	0.42958	-11.94378

(table continues)

	Predicted	Monte Carlo		Standard	Variance	Relative Bias	Relative Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
$\alpha_1 = 1.5, \alpha_2 = 2.0$, OVL = 0.830617							
100	0.00220538	0.815459	0.00238677	- 0.310276	1.0822461	- 1.82496	- 7.59958
500	0.00061676	0.835435	0.00062997	0.191967	1.0214133	0.58007	- 2.09644
$\alpha_1 = 2.5, \alpha_2 = 3.5$, OVL = 0.755917							
100	0.00258534	0.754192	0.00295801	- 0.031714	1.1441486	- 0.22818	-12.59877
500	0.00065256	0.763558	0.00061675	0.307690	0.9451338	1.01087	5.80513
$\alpha_1 = 1.5, \alpha_2 = 2.5$, OVL = 0.691640							
100	0.00276211	0.693917	0.00317429	0.040421	1.1492239	0.32927	-12.98475
500	0.00063438	0.703530	0.00060286	0.484261	0.9503242	1.71913	5.22725
$\alpha_1 = 2.5, \alpha_2 = 4.0$, OVL = 0.654885							
100	0.00279217	0.661026	0.00315809	0.109273	1.1310523	0.93769	-11.58676
500	0.00063021	0.668112	0.00059676	0.541438	0.9469238	2.01969	5.60512

(table continues)

						Relative	Relative
	Predicted	Monte Carlo		Standard	Variance	Bias	Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
$\alpha_1 = 2.0, \alpha_2 = 3.5$, OVL = 0.621790							
100	0.00278195	0.631951	0.00286496	0.189840	1.0298397	1.63419	- 2.89751
500	0.00061673	0.636882	0.00061714	0.607534	1.0006587	2.42727	- 0.06583
$\alpha_1 = 2.0, \alpha_2 = 4.0$, OVL = 0.529504							
100	0.00270272	0.542811	0.00267436	0.251306	0.9895098	2.51306	1.06014
500	0.00056157	0.546648	0.00059220	0.032377	1.0545355	3.23775	- 5.17152
$\alpha_1 = 1.5, \alpha_2 = 3.5$, OVL = 0.481225							
100	0.00255794	0.495070	0.00248876	0.277520	0.9729553	2.87698	2.77964
500	0.00054015	0.502054	0.00052604	0.908120	0.9738730	4.32818	2.68279
$\alpha_1 = 1.5, \alpha_2 = 4.0$, OVL = 0.401746							
100	0.00232771	0.419519	0.00222816	0.376528	0.9572334	4.42405	4.46773
500	0.00044278	0.423282	0.00043242	1.035634	0.9766022	5.36055	2.39541

Table C4

Results of the Monte Carlo Simulation Study: The Kernel Estimator of OVL Based on Independent Samples from Two Beta Distributions

						Relative	Relative
	Predicted	Monte Carlo		Standard	Variance	Bias	Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
$p_1 = 5, q_1 = 3, p_2 = 5$ and $\beta_1 = 3$, OVL = 1.000000							
100	0.0013893	0.892312	0.00097386	- 3.450789	0.7009629	-10.76879	42.66090
500	0.0002776	0.952482	0.00018356	- 3.507243	0.6613681	- 4.75180	51.20173
$p_1 = 2, q_1 = 2, p_2 = 3$ and $q_2 = 3$, OVL = 0.892669							
100	0.00174045	0.861193	0.00160425	- 0.785844	0.9217418	- 3.52599	8.49025
500	.000480867	0.890796	0.00053912	- 0.080651	1.1211597	- 0.20978	-10.80664
$p_1 = 2, q_1 = 2, p_2 = 1$ and $q_2 = 1$, OVL = 0.807550							
100	0.00207295	0.805789	0.00215627	- 0.379486	1.0401970	- 0.21821	- 3.86437
500	0.00068300	0.782637	0.00071247	- 0.933357	1.0431455	- 3.08504	- 4.13609
$p_1 = 5, q_1 = 3, p_2 = 3$ and $q_2 = 3$, OVL = 0.722479							
100	0.00277571	0.717282	0.00328779	- 0.090635	1.1844861	- 0.71932	-15.57520
500	0.00065956	0.730782	0.00068633	0.316390	1.0440814	1.14919	- 4.22203

Table C5

Results of the Monte Carlo Simulation Study: The Kernel Estimator of OVL Based on Independent Samples from a Standard Normal Distribution and a Standard Cauchy Distribution

						Relative	Relative
	Predicted	Monte Carlo		Standard	Variance	Bias	Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
				OVL = 0.748835			
100	0.00188767	0.731943	0.00197737	- 0.379879	1.0475164	-2.25581	-4.53610
500	0.00044291	0.751965	0.00045762	- 0.146309	1.0332051	0.41797	-3.21380

Table C6

Results of the Monte Carlo Simulation Study: The Kernel Estimator of OVL Based on Independent Samples from a Gamma Distribution with $\alpha = 3$ and a Chi Squared Distribution with 4 degrees of freedom

						Relative	Relative
	Predicted	Monte Carlo		Standard	Variance	Bias	Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
				OVL = 0.815890			
100	0.00235402	0.781535	0.00264487	- 0.668016	1.1235552	- 4.21073	-10.99681
500	0.00061259	0.795894	0.00063506	- 0.793482	1.0366929	- 2.45084	- 3.53942

APPENDIX D

MONTE CARLO PROGRAM FOR THE MAXIMUM LIKELIHOOD ESTIMATOR OF THE OVL

```
OPTIONS LS=132 PAGENO=1 NODATE;
LIBNAME SIM V604 'A:\';

DATA OUTPUT;
RUN;

DATA STAT;

ARRAY B{500} B1-B500;
ARRAY C{500} C1-C500;

N1=500;
N2=500;
M=1000;
SEED1=85652;
SEED2=32247;

PI=ARCOS(-1);

DO I=1 TO M;
DO J=LBOUND(B) TO HBOUND(B);
B(J) =0+SQRT(1)*NORMAL(SEED1);

END;

DO J = LBOUND(C) TO HBOUND(C);
C(J) = 1 + SQRT(1)*NORMAL(SEED2);
END;

MEAN1=MEAN(OF B1-B500);
MEAN2=MEAN(OF C1-C500);
VAR1B=VAR(OF B1-B500);
VAR2B=VAR(OF C1-C500);

VAR1=((VAR1B)*(N1-1))/N1;
VAR2=((VAR2B)*(N2-1))/N1;

KEEP MEAN1 MEAN2 VAR1 VAR2;



STD1=SQRT(VAR1);
STD2=SQRT(VAR2);

DELTAH = (MEAN1-MEAN2)/STD1;
```

```
GAMMAH=VAR2/VAR1;

A = SQRT(DELTAH**2 + (GAMMAH-1)*LOG(GAMMAH));

Z11 = (DELTAH - SQRT(GAMMAH)*A)/(GAMMAH-1);
Z22 = (SQRT(GAMMAH)*DELTAH + A)/(GAMMAH-1);
Z12 = (SQRT(GAMMAH)*DELTAH - A)/(GAMMAH-1);
Z21 = (DELTAH + SQRT(GAMMAH)*A)/(GAMMAH-1);

PHIZ11 = PROBNORM(Z11);
PHIZ22 = PROBNORM(Z22);
PHIZ12 = PROBNORM(Z12);
PHIZ21 = PROBNORM(Z21);

OVL = PHIZ11 + PHIZ22 - PHIZ12 - PHIZ21 +1;

DZ11D = (1 - SQRT(GAMMAH)*DELTAH*A**(-1))/(GAMMAH-1);
DZ22D = (SQRT(GAMMAH)+DELTAH*A**(-1))/(GAMMAH-1);
DZ12D = (SQRT(GAMMAH)-DELTAH*A**(-1))/(GAMMAH-1);
DZ21D = (1 + SQRT(GAMMAH)*DELTAH*A**(-1))/(GAMMAH-1);

DZ11G = ((2*SQRT(GAMMAH)*A - 2*DELTAH) -
(GAMMAH**(-1/2)*A*(GAMMAH-1)+ SQRT(GAMMAH)*A**(-1)*
(((GAMMAH-1)/GAMMAH) + LOG(GAMMAH))*(GAMMAH-1)))/(2*(GAMMAH
-1)**2);

DZ22G = ((GAMMAH**(-1/2)*(GAMMAH-1)*DELTAH +
A**(-1)*(((GAMMAH-1)/GAMMAH) + LOG(GAMMAH))*
(GAMMAH-1)) - (2*SQRT(GAMMAH)*DELTAH + 2 * A))/(2*(GAMMAH -1)**2);

DZ12G = ((GAMMAH**(-1/2)*(GAMMAH-1)*DELTAH -
A**(-1)*(((GAMMAH-1)/GAMMAH) + LOG(GAMMAH))*
(GAMMAH-1)) - (2*SQRT(GAMMAH)*DELTAH - 2 * A))/(2*(GAMMAH -1)**2);

DZ21G = ((-2*SQRT(GAMMAH)*A - 2*DELTAH) +
(GAMMAH**(-1/2)*A*(GAMMAH-1)+ SQRT(GAMMAH)*A**(-1)*
(((GAMMAH-1)/GAMMAH) + LOG(GAMMAH))*(GAMMAH-1)))/(2*(GAMMAH
-1)**2);

VARG = (2*N1**2*(N2-1)*(N1+N2-4))/(N2**2*(N1-3)**2*(N1-5))*GAMMAH**2;

Z1=(N1-2)/2;
Z2=(N1-1)/2;

GAMZ11 = EXP(-Z1/3)*EXP(-Z1/3)*EXP(-Z1/3);
```

```
GAMZ2A = Z1**((Z1-(1/2))/3);

GAMZ2B = Z1**((Z1-(1/2))/3);

GAMZ2C = Z1**((Z1-(1/2))/3);

GAMZ3 = (2*3.141592654)**(1/2);

GAMZ4 = (1 + 1/(12*Z1) + 1/(288*Z1**2) - 139/(51840*Z1**3) -
571/(2488320*Z1**4));

GAMZ12 = EXP(-Z2/3)*EXP(-Z2/3)*EXP(-Z2/3);

GAMZ22A= Z2**((Z2-(1/2))/3);

GAMZ22B= Z2**((Z2-(1/2))/3);

GAMZ22C= Z2**((Z2-(1/2))/3);

GAMZ32 = (2*3.141592654)**(1/2);

GAMZ42 = (1 + 1/(12*Z2) + 1/(288*Z2**2) - 139/(51840*Z2**3) -
571/(2488320*Z2**4));

GAM1=GAMZ11/GAMZ12;

GAM2A = GAMZ2A/GAMZ22A;

GAM2B = GAMZ2B/GAMZ22B;

GAM2C = GAMZ2C/GAMZ22C;

GAM3 = GAMZ3/GAMZ32;

GAM4 = GAMZ4/GAMZ42;

GAM = GAM1*GAM2A*GAM2B*GAM2C*GAM3*GAM4;

VARD = (1 + GAMMAH*(N1/N2))/(N1-3) + DELTAH**2 *(N1/(N1-3) -
((SQRT(N1/2)*(GAM))**2));

Z1=(N1-4)/2;
Z2=(N1-1)/2;

GAMZ11 = EXP(-Z1/3)*EXP(-Z1/3)*EXP(-Z1/3);
```

```
GAMZ2A = Z1**((Z1-(1/2))/3);

GAMZ2B = Z1**((Z1-(1/2))/3);

GAMZ2C = Z1**((Z1-(1/2))/3);

GAMZ3 = (2*3.141592654)**(1/2);

GAMZ4 = (1 + 1/(12*Z1) + 1/(288*Z1**2) - 139/(51840*Z1**3) -
571/(2488320*Z1**4));

GAMZ12 = EXP(-Z2/3)*EXP(-Z2/3)*EXP(-Z2/3);

GAMZ22A= Z2**((Z2-(1/2))/3);

GAMZ22B= Z2**((Z2-(1/2))/3);

GAMZ22C= Z2**((Z2-(1/2))/3);

GAMZ32 = (2*3.141592654)**(1/2);

GAMZ42 = (1 + 1/(12*Z2) + 1/(288*Z2**2) - 139/(51840*Z2**3) -
571/(2488320*Z2**4));

GAM1=GAMZ11/GAMZ12;

GAM2A = GAMZ2A/GAMZ22A;

GAM2B = GAMZ2B/GAMZ22B;

GAM2C = GAMZ2C/GAMZ22C;

GAM3 = GAMZ3/GAMZ32;

GAM4 = GAMZ4/GAMZ42;

GAM2 = GAM1*GAM2A*GAM2B*GAM2C*GAM3*GAM4;

COVDG = DELTAH*GAMMAH*(N2-1)/N2*((N1/2)**(3/2)*(GAM2))*(1/(N1-3));
PHZ11=(EXP(-Z11**2/2))/SQRT(2*PI);
PHZ12=(EXP(-Z12**2/2))/SQRT(2*PI);
PHZ21=(EXP(-Z21**2/2))/SQRT(2*PI);
PHZ22=(EXP(-Z22**2/2))/SQRT(2*PI);

VAROVLA = (PHZ11*DZ11D + PHZ22*DZ22D - PHZ12*DZ12D - PHZ21*DZ21D);
```

```
VAROVLB = (PHZ11*DZ11G + PHZ22*DZ22G - PHZ12*DZ12G - PHZ21*DZ21G);
VAROVLC = VAROVLA*VAROVLB;

VAROVL = VAROVLA**2*VARD + VAROVLB**2*VARG +
2*VAROVLC*COVDG;

KEEP  J GAMMAH DELTAH OVL VAROVL;
OUTPUT; END;
PROC MEANS NOPRINT MEAN VAR ; VAR OVL VAROVL; OUTPUT
OUT=STAT2 MEAN=MCOVL PREDVAR VAR=MCOVLVAR VARVAR;

DATA STAT1;

DELTAH = 0;
GAMMAH=1;
N1=500;
N2=500;

A = SQRT(DELTAH**2 + (GAMMAH-1)*LOG(GAMMAH));

Z11 = (DELTAH - SQRT(GAMMAH)*A)/(GAMMAH-1);
Z22 = (SQRT(GAMMAH)*DELTAH + A)/(GAMMAH-1);
Z12 = (SQRT(GAMMAH)*DELTAH - A)/(GAMMAH-1);
Z21 = (DELTAH + SQRT(GAMMAH)*A)/(GAMMAH-1);

PHIZ11 = PROBNORM(Z11);
PHIZ22 = PROBNORM(Z22);
PHIZ12 = PROBNORM(Z12);
PHIZ21 = PROBNORM(Z21);

TOVL = PHIZ11 + PHIZ22 - PHIZ12 - PHIZ21 +1;


KEEP TOVL N1 DELTAH GAMMAH;


DATA A;
MERGE STAT1 STAT2;

KEEP MCOVL MCOVLVAR TOVL PREDVAR N1 DELTAH GAMMAH;


DATA STAT3;
SET A;
```

```
STDBS = (MCOVL - TOVL)/(SQRT(MCOVLVAR));
VARATIO = (PREDVAR/MCOVLVAR);

STDBSV = 1 - VARATIO;

RLBIASOL = (MCOVL-TOVL)/TOVL;
RLBIASV = (PREDVAR-MCOVLVAR)/MCOVLVAR;

DATA OUTPUT; SET OUTPUT STAT3; RUN;


DATA OUTPUT; SET OUTPUT; IF N1=. THEN DELETE; RUN;
PROC APPEND BASE=SIM.OUTPUT DATA=OUTPUT;
RUN;
```

APPENDIX E

RESULTS OF THE MONTE CARLO SIMULATION STUDY: MAXIMUM LIKELIHOOD ESTIMATOR OF OVL

Table E1

Results of the Monte Carlo Simulation Study: Maximum-Likelihood Estimator of OVL Based on Independent Samples from Two Normal Distributions

N	Predicted Variance	Monte Carlo Mean	Monte Carlo Variance	Standard Bias	Variance Ratio	Relative Bias OVL (%)	Relative Bias Variance (%)
$\mu_1 = 2, \sigma^2_1 = 4, \mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.808473							
100	0.00300140	0.805266	0.00282875	- 0.06031	1.0610348	- 0.39670	6.10348
500	0.00060004	0.807427	0.00062293	- 0.04192	0.9632467	- 0.12941	- 3.67534
$\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 0$ and $\sigma^2_2 = 3$, OVL = 0.740641							
100	0.00217388	0.737767	0.00193680	- 0.06532	1.1224109	- 0.38813	12.24110
500	0.00040968	0.739943	0.00039606	- 0.03510	1.0343996	- 0.09431	3.43996
$\mu_1 = 5, \sigma^2_1 = 10, \mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.683020							
100	0.00258851	0.674511	0.00261027	- 0.16654	0.9916621	- 1.24575	- 0.83379
500	0.00050871	0.681301	0.00050871	- 0.07541	0.9790895	- 0.25165	- 2.09105
$\mu_1 = 1, \sigma^2_1 = 1, \mu_2 = 0$ and $\sigma^2_2 = 3$, OVL = 0.639430							
100	0.00215594	0.631886	0.00195512	- 0.17061	1.1027150	- 1.17980	10.27150
500	0.00041049	0.637845	0.00043590	- 0.07591	0.9417093	- 0.24786	- 5.82908

(table continues)

						Relative	Relative
	Predicted	Monte Carlo		Standard	Variance	Bias	Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
$\mu_1 = 1, \sigma^2_1 = 1, \mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 0.609934							
100	0.00189284	0.607985	0.00179963	- 0.04594	1.0517945	- 0.31955	5.17945
500	0.00035856	0.609155	0.00033517	- 0.04259	1.0697860	- 0.12784	6.97861
$\mu_1 = 0, \sigma^2_1 = 3, \mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 0.588750							
100	0.00275533	0.587354	0.00262018	- 0.02721	1.0515834	- 0.23710	5.15834
500	0.00054066	0.587747	0.00056857	- 0.04206	0.9509089	- 0.17035	- 4.90911
$\mu_1 = 1, \sigma^2_1 = 1, \mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.457402							
100	0.00169228	0.455049	0.00162205	- 0.05843	1.0432971	- 0.51445	4.32971
500	0.00032703	0.456442	0.00032307	- 0.20982	1.0122758	- 0.20982	1.22759
$\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 0.453388							
100	0.00189513	0.450932	0.00181286	- 0.05768	1.0453843	- 0.54170	4.53843
500	0.00037053	0.452333	0.00037433	- 0.05452	0.9898433	- 0.23266	- 1.01567

(table continues)

						Relative	Relative
	Predicted	Monte Carlo		Standard	Variance	Bias	Bias
N	Variance	Mean	Variance	Bias	Ratio	OVL (%)	Variance (%)
			$\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.315318				
100	0.00152879	0.313225	0.00145376	- 0.05490	1.0516109	- 0.66387	5.16109
500	0.00030141	0.314365	0.00031082	- 0.30218	0.9697246	- 0.30218	- 3.02754

APPENDIX F

STANDARD BIAS AND RELATIVE INEFFICIENCY OF THE KERNEL ESTIMATOR OF OVL

Table F1

Standard Bias and Relative Inefficiency of the Kernel Estimator of the OVL for Comparison to the Maximum-Likelihood Estimator of the OVL

N	Standard Bias	Relative Inefficiency
$\mu_1 = 2$, $\sigma^2_1 = 4$, $\mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.808473		
100	0.2018649	1.0171030
500	0.1202793	1.0819996
$\mu_1 = 0$, $\sigma^2_1 = 1$, $\mu_2 = 0$ and $\sigma^2_2 = 3$, OVL = 0.740641		
100	-0.3670605	1.3152623
500	-0.2169712	1.4490980
$\mu_1 = 5$, $\sigma^2_1 = 10$, $\mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.683020		
100	-0.1678776	1.1074716
500	0.2567309	1.1343393
$\mu_1 = 1$, $\sigma^2_1 = 1$, $\mu_2 = 0$ and $\sigma^2_2 = 3$, OVL = 0.639430		
100	-0.1710212	1.2956545
500	0.2216664	1.2207387
$\mu_1 = 1$, $\sigma^2_1 = 1$, $\mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 0.609934		
100	-0.1648907	1.3246945
500	0.1958744	1.6166125
$\mu_1 = 1$, $\sigma^2_1 = 1$, $\mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.457402		
100	0.2018649	1.1293575
500	0.5983299	1.0046608
$\mu_1 = 1$, $\sigma^2_1 = 1$, $\mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.457402		
100	0.0704164	1.3953577
500	0.6022539	1.3439440

(table continues)

N	Standard Bias	Relative Inefficiency
$\mu_1 = 0$, $\sigma^2_1 = 1$, $\mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 0.453388		
100	0.1425629	1.1835608
500	0.6994146	1.1914087
$\mu_1 = 1$, $\sigma^2_1 = 1$, $\mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.457402		
100	0.2689872	1.2247620
50	0.9796321	1.2074191

APPENDIX G

MODELING OF THE BIAS OF THE KERNEL ESTIMATOR OF THE OVL

Table G1

Model: Two Normal Distributions

Source	df	SS	MS	F-Value
N	1	0.00279811	0.00279811	7.89*
MCOVL^2	1	0.04213873	0.04213873	120.14**
MCOVL^3	1	0.00652833	0.00652833	18.61**
Error	20	0.00701516	0.00035076	
Total	23	0.94532505		
* $p < 0.05$	**$p < 0.01$			

Rsquare = 0.992579

Parameter Estimates

	Parameter		T for HO:
Variable	Estimate	Standard Error	Parameter = 0
Intercept	0.208030872	0.02053459	10.13**
N(Sample Size)	- 0.000054292	0.00001922	- 2.82*
MCOVL^2	1.426573818	0.13015396	10.96**
MCOVL^3	- 0.530712853	0.12301619	- 4.31**
* $p < 0.05$	**$p < 0.01$		

Table G2

Predicted Values of the OVL for the Model for reducing the Bias of the Kernel Estimator of the OVL for two Normal Distributions

				MODEL	OLD	
N	OVL	MCOVL	YHAT	RELATIVE	RELATIVE	PERCENT
				BIAS %	BIAS %	CHANGE
100	0.31532	0.32557	0.33550	6.401	3.253	- 3.148
500	0.31532	0.33259	0.31916	1.219	5.477	4.259
100	0.45339	0.45946	0.45228	- 0.245	1.339	1.094
500	0.45339	0.46692	0.43787	- 3.422	2.985	- 0.437
100	0.45740	0.46024	0.45304	- 0.954	0.620	- 0.334
500	0.45740	0.46823	0.43916	- 3.990	2.367	- 1.621
100	0.58875	0.59908	0.60049	1.994	1.755	- 0.239
500	0.58875	0.60302	0.58326	- 0.933	2.423	1.490
100	0.60993	0.60294	0.60488	- 0.828	- 1.147	0.319
500	0.60993	0.61352	0.59530	- 2.400	0.588	- 1.812
100	0.61708	0.62458	0.62980	2.062	1.216	- 0.846
500	0.61708	0.63083	0.61535	- 0.279	2.229	1.950
100	0.63943	0.63187	0.63828	- 0.179	- 1.183	1.004
500	0.63943	0.64406	0.63086	- 1.341	0.072	- 0.617
100	0.68302	0.67444	0.68870	0.831	- 1.256	0.425
500	0.68302	0.68887	0.68437	0.197	0.857	0.660
100	0.74064	0.72449	0.74957	1.206	- 2.181	0.976
500	0.74064	0.73632	0.74246	0.246	- 0.583	0.337
100	0.80847	0.79634	0.83927	3.809	- 1.500	- 2.308
500	0.80847	0.81148	0.83669	3.490	0.300	- 3.118
100	1.00000	0.89157	0.96046	- 3.954	-10.843	6.889
100	1.00000	0.89212	0.96116	- 3.884	-10.788	6.904
500	1.00000	0.95105	1.01469	1.469	- 4.895	3.426
500	1.00000	0.95072	1.01426	1.426	- 4.928	3.884

Table G3

Model: Two Weibull Distributions

Source	df	SS	MS	F Value
N	1	0.00483966	0.00483966	7.45*
MCOVL^2	1	0.60604026	0.60604026	932.78**
Error	21	0.01364394	0.00064971	
Total	23	0.61968420		

* $p < 0.05$ **$p < 0.01$

Rsquare = 0.977982

Parameter Estimates

Variable	Parameter Estimate	Standard Error	T for HO: Parameter = 0
Intercept	0.3310906562	0.01696710	19.51**
N(Sample Size)	- 0.0000712874	0.00002612	- 2.73*
MCOVL^2	0.7794010304	0.02551940	30.54**

* $p < 0.05$ **$p < 0.01$

Table G4

Predicted Values of the OVL for the Model for reducing the Bias of the Kernel Estimator of the OVL for Two Weibull Distributions

				MODEL	OLD	
N	OVL	MCOVL	YHAT	RELATIVE	RELATIVE	PERCENT
				BIAS %	BIAS %	CHANGE
100	0.42141	0.39940	0.44829	6.378	- 5.223	- 1.155
500	0.42141	0.40422	0.42280	0.328	- 4.080	3.751
100	0.57323	0.58125	0.58728	2.451	1.399	- 1.053
500	0.57323	0.58713	0.56412	- 1.590	2.424	0.835
100	0.64210	0.62490	0.62832	- 2.146	- 2.678	0.532
500	0.64210	0.63330	0.60804	- 5.304	- 1.370	- 3.934
100	0.69098	0.69229	0.69750	0.944	0.190	- 0.754
500	0.69098	0.69970	0.67702	- 2.020	1.262	- 0.758
100	0.73725	0.72512	0.73377	- 0.472	- 1.646	1.173
500	0.73725	0.73591	0.71754	- 2.673	- 0.182	- 2.491
100	0.74739	0.73890	0.74949	0.281	- 1.136	0.855
500	0.74739	0.75122	0.73529	- 1.619	0.513	- 1.106
100	0.80280	0.79746	0.81961	2.094	0.665	- 1.429
500	0.80280	0.81577	0.81413	1.411	1.616	0.205
100	0.81490	0.79696	0.81899	0.502	- 2.202	1.699
500	0.81490	0.81713	0.81585	0.117	0.273	0.157
100	0.85697	0.85624	0.89538	4.482	- 0.085	- 4.398
500	0.85697	0.88831	0.91047	6.243	3.657	- 2.586
100	0.86784	0.84429	0.87954	1.348	- 2.714	1.366
500	0.86784	0.86847	0.88330	1.781	0.072	- 1.709
100	1.00000	0.89558	0.94909	- 5.091	-10.442	5.351
100	1.00000	0.89451	0.94760	- 5.240	-10.549	5.309
500	1.00000	0.95331	1.00377	0.377	- 4.669	4.292
500	1.00000	0.95248	1.00254	0.254	- 4.752	4.498

Table G5

Model: Two Gamma Distributions

Source	df	SS	MS	F-Value
N	1	0.00463796	0.00463796	9.94**
MCOVL^2	1	0.84338585	0.84338585	1806.76**
Error	21	0.00980266	0.00046679	
Total	23	0.85318851		
* $p < 0.05$	**$p < 0.01$			

Rsquare = 0.988511

Parameter Estimates

Variable	Parameter Estimate	Standard Error	T for HO: Parameter = 0
Intercept	0.2963542028	0.01286178	23.04**
N(Sample Size)	- 0.0000696988	0.00002211	- 3.15**
MCOVL^2	0.8321760157	0.01957782	42.51**
* $p < 0.05$	**$p < 0.01$		

Table G6

Predicted Values of the OVL for the Model for Reducing the Bias of the Kernel Estimator of the OVL for Two Gamma Distributions

				MODEL	OLD	
				RELATIVE	RELATIVE	PERCENT
N	OVL	MCOVL	YHAT	BIAS %	BIAS %	CHANGE
100	0.40175	0.41952	0.43584	8.488	4.424	- 4.064
500	0.40175	0.42328	0.41060	2.205	5.361	3.156
100	0.48123	0.49507	0.49335	2.519	2.877	0.358
500	0.48123	0.50205	0.47126	- 2.070	4.328	2.258
100	0.52950	0.54281	0.53458	0.959	2.513	1.555
500	0.52950	0.54665	0.51018	- 3.650	3.238	- 0.412
100	0.62179	0.63195	0.62172	- 0.011	1.634	1.624
500	0.62179	0.63688	0.59905	- 3.657	2.427	- 1.230
100	0.65489	0.66103	0.65301	- 0.287	0.938	0.651
500	0.65489	0.66811	0.63297	- 3.347	2.020	- 1.327
100	0.69164	0.69392	0.69009	- 0.224	0.329	0.106
500	0.69164	0.70353	0.67339	- 2.638	1.719	- 0.919
100	0.75592	0.75419	0.76273	0.904	- 0.228	- 0.673
500	0.75592	0.76356	0.74668	- 1.222	1.011	- 0.211
100	0.83062	0.81546	0.84276	1.462	- 1.825	0.363
500	0.83062	0.83544	0.84232	1.409	0.580	- 0.829
100	0.85448	0.83381	0.86794	1.576	- 2.419	0.844
500	0.85448	0.85815	0.87434	2.324	0.430	- 1.894
100	0.89132	0.85786	0.90180	1.176	- 3.754	2.578
500	0.89132	0.89002	0.92071	3.297	- 0.146	- 3.151
100	1.00000	0.89409	0.95462	- 4.534	-10.591	6.053
100	1.00000	0.89412	0.95467	- 4.533	-10.588	6.055
500	1.00000	0.95250	1.01651	1.651	- 4.750	3.099
500	1.00000	0.95163	1.01512	1.512	- 4.837	3.326

APPENDIX H

RESULTS OF THE MONTE CARLO SIMULATION STUDY: KERNEL ESTIMATOR OF THE OVL USING THE ALTERNATIVE REFERENCE RULE

Table H1

Results of the Monte Carlo Simulation Study: The Kernel Estimator of the OVL Using The Alternative Normal Reference Rule

				Relative
	Monte Carlo		Standard	Bias
N	Mean	Variance	Bias	OVL (%)
$\mu_1 = 0$, $\sigma^2_1 = 1$, $\mu_2 = 0$ and $\sigma^2_2 = 1$, OVL = 1.000000				
100	0.936678	0.00071107	- 2.37458	- 6.33202
500	0.971878	0.00012386	- 2.52692	- 2.81223
$\mu_1 = 2$, $\sigma^2_1 = 4$, $\mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 1.000000				
100	0.935638	0.00069486	- 2.44163	- 6.43620
500	0.971051	0.00015674	- 2.31229	- 2.89488
$\mu_1 = 2$, $\sigma^2_1 = 4$, $\mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.808473				
100	0.851653	0.00142078	1.14557	5.34096
500	0.862247	0.00030772	3.08129	6.68568
$\mu_1 = 0$, $\sigma^2_1 = 1$, $\mu_2 = 0$ and $\sigma^2_2 = 3$, OVL = 0.740641				
100	0.725801	0.00225026	- 0.31284	- 2.00371
500	0.728332	0.00044842	- 0.58129	- 1.66198
$\mu_1 = 5$, $\sigma^2_1 = 10$, $\mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.683020				
100	0.742837	0.00165764	1.46920	8.75774
500	0.751072	0.00033362	3.72580	9.96348
$\mu_1 = 1$, $\sigma^2_1 = 1$, $\mu_2 = 0$ and $\sigma^2_2 = 3$, OVL = 0.639430				
100	0.677575	0.00166378	0.93518	5.96553
500	0.681896	0.00034403	2.28953	6.64133

(table continues)

				Relative
	Monte Carlo		Standard	Bias
N	Mean	Variance	Bias	OVL (%)
$\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 1$ and $\sigma^2_2 = 1$, OVL = 0.617075				
100	0.730267	0.00158729	2.84110	18.34330
500	0.732179	0.00026977	8.86412	18.65320
$\mu_1 = 1, \sigma^2_1 = 1, \mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 0.609934				
100	0.632031	0.00165720	0.54280	3.62276
500	0.632441	0.00031625	1.26562	3.69006
$\mu_1 = 0, \sigma^2_1 = 3, \mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 0.588750				
100	0.703841	0.00149100	2.98060	19.54842
500	0.708157	0.00030745	6.59874	20.281578
$\mu_1 = 1, \sigma^2_1 = 1, \mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.457402				
100	0.528375	0.00118608	2.06081	15.51662
500	0.533580	0.00024381	4.87868	16.65445
$\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 2$ and $\sigma^2_2 = 4$, OVL = 0.453388				
100	0.547419	0.00119628	2.71865	20.73957
500	0.551347	0.00024535	6.25392	21.60599
$\mu_1 = 0, \sigma^2_1 = 1, \mu_2 = 3$ and $\sigma^2_2 = 5$, OVL = 0.315318				
100	0.440854	0.00101119	3.94792	39.81402
500	0.442932	0.00020726	8.86412	40.47132

APPENDIX I

DATA USED FOR EXAMPLES OF THE KERNEL ESTIMATOR OF THE OVL

Table 1

Natural Logarithm of Estimated Wealth ($) of Alabama Farm Operators in 1850

Farmers Who Persisted to 1860 (N = 317)

4.21416	4.21416	4.25323	4.56381	4.64991	4.73280	4.90533	5.07689	5.11912	5.20518
5.40509	5.40509	5.42741	5.44254	5.44473	5.45063	5.49191	5.53405	5.53529	5.53899
5.60263	5.60263	5.60400	5.64703	5.69217	5.73735	5.74271	5.77358	5.77829	5.80419
5.82005	5.82005	5.82648	5.83656	5.84276	5.84852	5.88618	5.89659	5.90832	5.90980
5.94187	5.94187	5.97209	5.97529	5.98740	6.00174	6.01016	6.02725	6.03722	6.04100
6.05349	6.05349	6.06085	6.08041	6.10489	6.11453	6.17563	6.18173	6.18475	6.19665
6.20839	6.20839	6.22106	6.22539	6.22588	6.23464	6.25085	6.25871	6.25941	6.26258
6.29788	6.29788	6.30155	6.30818	6.32650	6.33811	6.34261	6.34819	6.35085	6.35309
6.36389	6.36389	6.38295	6.39990	6.40032	6.40186	6.41668	6.42937	6.44204	6.45252
6.46440	6.46440	6.46913	6.48949	6.49052	6.49404	6.52893	6.56313	6.56939	6.58554
6.62493	6.62493	6.63224	6.63248	6.64069	6.65178	.68081	6.68311	6.68788	6.69689
6.70338	6.70338	6.72479	6.74065	6.74329	6.75940	6.76556	6.79926	6.80825	6.81386
6.84763	6.84763	6.86236	6.89961	6.91177	6.92975	6.93548	6.93619	6.93809	6.94488

(table continues)

Farmers Who Persisted to 1860 (N = 317)									
6.94724	6.94724	6.97980	6.89770	6.99769	7.04210	7.07712	7.09184	7.09526	7.10931
7.14166	7.14166	7.16208	7.16836	7.17715	7.17718	7.20558	7.21872	7.24835	7.25009
7.26425	7.26425	7.26760	7.26927	7.28263	7.28912	7.30172	7.30367	7.34225	7.37290
7.40502	7.40502	7.41431	7.43926	7.46007	7.46746	7.47636	7.48269	7.51404	7.51766
7.33563	7.53363	7.53483	7.54618	7.57841	7.59655	7.59657	7.60043	7.63358	7.65087
7.71335	7.71335	7.78054	7.78825	7.79117	7.79370	7.82290	7.86128	7.87158	7.92097
7.95330	7.95330	7.98781	8.00523	8.00786	8.01839	8.03306	8.04002	8.06254	8.06639
8.08577	8.08577	8.09366	8.09549	8.12507	8.13436	8.17189	8.17353	8.20650	8.22955
8.25031	8.25031	8.26887	8.31734	8.33159	8.34445	8.35271	8.43234	8.44740	8.44750
8.46904	8.46904	8.48987	8.51567	8.51695	8.56893	8.65572	8.66603	8.67946	8.68544
8.69551	8.69551	8.70941	8.72426	8.73835	8.80387	8.81081	8.86700	8.87300	8.87992
8.90051	8.90051	8.93037	8.94784	8.95383	8.96772	8.97551	8.97623	8.98091	8.98935
9.00785	9.00785	9.02316	9.04473	9.05862	9.08744	9.11221	9.11630	9.11726	9.11835
9.20084	9.20084	9.20626	9.24920	9.26906	9.27423	9.29117	9.31422	9.35204	9.35661

(table continues)

Farmers Who Persisted to 1860 (N = 317)

9.38209	9.38209	9.39970	9.40017	9.44978	9.47143	9.47947	9.48789	9.61620	9.63652
9.69919	9.69919	9.72719	9.73754	9.74478	9.74801	9.75506	9.77395	9.79344	9.83127
9.84009	9.84009	9.85847	9.87027	9.94377	9.95733	9.99520	10.01175	10.04653	10.06632
10.11141	10.11141	10.12518	10.16019	10.17805	10.22034	10.26689	10.29022	10.45151	10.55801
10.68460	10.68460	10.78461	10.84595	10.85365	10.92429	11.09359			

Farmers Who Did Not Persist to 1869 (N=284)

3.22865	3.22865	3.34510	3.47189	3.81473	3.93256	4.21257	4.24103	4.26971	4.40593
4.48537	4.48537	4.53567	4.62215	4.92249	5.01930	5.08780	5.17036	5.17768	5.18133
5.22241	5.22241	5.22378	5.22744	5.22862	5.26090	5.26587	5.27674	5.28179	5.30354
5.35598	5.35598	5.38269	5.42761	5.46931	5.47113	5.49400	5.50709	5.52812	5.53407
5.55836	5.55836	5.56169	5.59500	5.60516	5.65823	5.67502	5.69517	5.69685	5.70189
5.70704	5.70704	5.71368	5.71909	5.77522	5.78250	5.78418	5.80651	5.81317	5.84040
5.87017	5.87017	5.93218	5.93701	5.95070	5.95475	5.95720	5.96532	5.98925	5.99823

(table continues)

Farmers Who Did Not Persisted to 1860 (N = 284)

6.01106	6.01106	6.02962	6.03436	6.06322	6.07099	6.07389	6.07407	6.07879	6.07980
6.14296	6.14296	6.15709	6.19372	6.19533	6.19727	6.19835	6.21624	6.21956	6.24181
6.27650	6.27650	6.27678	6.30243	6.33577	6.33779	6.35118	6.35201	6.36225	6.36611
6.37515	6.37515	6.38342	6.40162	6.40429	6.40639	6.42582	6.45425	6.47854	6.48077
6.48698	6.48698	6.48843	6.49554	6.51322	6.52731	6.53742	6.55105	6.55262	6.55336
6.57088	6.57088	6.58412	6.58991	6.59563	6.59720	6.60166	6.62030	6.65888	6.67885
6.70185	6.70185	6.70888	6.71655	6.72904	6.73345	6.76300	6.76888	6.78850	6.78925
6.80181	6.80181	6.83908	6.85027	6.855985	6.85606	6.87894	6.92208	6.94465	6.94826
6.96306	6.96306	6.97092	6.97313	7.00796	7.01590	7.02235	7.04753	7.05803	7.06198
7.10254	7.10254	7.11291	7.15667	7.16115	7.16659	7.17223	7.18402	7.21773	7.22309
7.24344	7.24344	7.26708	7.28546	7.31530	7.31923	7.32451	7.34625	7.35535	7.38495
7.43854	7.43854	7.48142	7.49396	7.49503	7.52466	7.54876	7.55728	7.62239	7.63052
7.64660	7.64660	7.64701	7.68918	7.72842	7.73653	7.73838	7.77271	7.81490	7.82789
7.86557	7.86557	7.88472	7.89815	7.94927	7.95869	7.99312	8.01199	8.02907	8.07602
8.09628	8.09628	8.11817	8.12403	8.13362	8.15256	8.18828	8.20372	8.20553	8.21440

(table continues)

Farmers Who Persisted to 1860 (N = 317)

8.26488	8.26488	8.26591	8.31669	8.31909	8.37300	8.39874	8.40818	8.42732	8.46627
8.49997	8.49997	8.53793	8.56129	8.66589	8.69371	8.71659	8.72214	8.75199	8.75490
8.83356	8.83356	8.84830	8.86652	8.88433	8.88945	8.90112	8.91990	8.92617	8.98861
9.00271	9.00271	9.01954	9.02716	9.11180	9.15928	9.16358	9.18848	9.19234	9.21431
9.26696	9.26696	9.35905	9.38043	9.38186	9.49479	9.59331	9.60459	6.63225	9.69350
9.78480	9.78480	9.78553	9.84848	9.93786	9.98421	9.98744	10.21750	10.23677	10.34977
10.95965	10.95965	11.26462	11.47963						

Table 2

Irish Educational Transition Data: Drumcondra Verbal Reasonng Test Score

Male students' Drumcondra Verbal Reasoning Test scores (N = 231)									
71	72	72	73	74	76	78	78	79	79
80	80	81	81	81	82	82	83	83	84
84	84	84	84	84	85	85	85	86	86
86	86	86	87	87	87	88	88	88	88
88	89	89	90	90	90	90	90	90	90
90	90	91	91	91	91	92	92	92	92
93	93	93	93	94	94	94	95	95	95
95	95	96	96	96	96	97	97	97	98
98	98	98	99	99	99	99	99	99	99
99	100	100	100	100	101	101	101	101	101
101	101	102	102	102	102	103	103	103	103
103	103	103	103	103	104	104	104	104	104
104	104	105	105	106	106	106	106	106	107
107	107	107	107	107	107	108	108	108	108
108	108	108	109	109	109	109	109	109	109
110	110	110	110	110	111	112	112	112	112
113	113	113	113	113	113	113	114	114	114
114	114	114	114	114	115	115	115	116	116
116	117	117	117	117	118	118	118	119	119
120	120	121	121	121	121	122	122	122	123
123	123	123	123	124	124	124	125	125	125
125	125	126	126	126	127	127	127	127	129
129	129	130	131	132	134	135	136	136	137
140									

(table continues)

Female students' Drumcondra Verbal Reasoning Test scores (N = 238)

72	72	73	73	75	75	77	78	78	81
81	81	81	82	82	83	83	84	84	84
84	85	85	85	85	85	85	85	86	86
86	86	86	86	87	87	88	88	88	89
89	89	89	90	90	90	90	90	90	90
90	90	91	91	91	91	91	91	91	91
91	92	92	92	92	92	93	93	93	93
93	93	93	93	93	94	94	94	94	94
94	94	94	94	94	95	96	96	96	97
97	97	97	97	97	97	98	98	98	99
99	99	99	99	99	99	100	100	100	100
100	100	100	100	100	100	100	100	100	101
101	101	101	101	102	102	102	102	102	102
102	102	102	103	103	103	103	103	103	103
103	104	104	104	104	104	104	104	104	104
104	104	104	105	105	105	105	105	105	106
106	106	106	106	106	107	107	107	107	107
108	108	108	108	109	109	109	109	109	109
109	109	109	109	110	110	110	110	111	111
111	111	111	111	111	111	112	112	112	113
113	114	114	114	114	114	114	114	114	115
115	115	115	116	116	116	117	117	117	117
118	118	119	119	120	120	120	122	122	123
123	123	123	127	127	127	134	135		

Table 3

Acute Myocardial Infarction Registry: Minutes from onset of ischemic chest pain to ECG

Male Patients who experienced chest pain for more than
6 hours before treatment (N = 131)

17	17	51	58	71	78	88	100	117	180
205	251	253	297	316	321	325	333	333	343
350	355	375	387	389	394	394	395	397	403
407	416	416	423	425	425	431	432	443	444
446	450	450	453	455	459	461	465	470	471
471	474	474	478	478	480	480	485	490	492
493	497	499	500	503	508	512	513	517	520
521	525	531	538	540	549	550	559	562	564
568	569	570	576	576	577	577	577	582	597
599	603	616	617	624	624	626	630	634	636
637	641	648	649	655	659	667	670	674	675
676	692	696	700	702	712	720	720	725	728
733	751	762	772	800	808	840	900	1305	1377
1435									

Female Patients who experienced chest pain for more than 6 hours before treatment
(N = 69)

115	213	258	303	370	380	386	401	408	414
434	435	443	446	450	472	473	475	476	491
495	501	505	505	514	516	522	530	530	531
535	544	549	556	565	566	570	570	570	571
572	577	583	587	590	595	604	617	633	635
635	661	664	684	689	692	693	696	697	700
711	716	717	720	729	745	750	769	870	

APPENDIX J

RESULTS OF THE EXAMPLES OF THE KERNEL ESTIMATOR OF THE OVL

	Example 1	Example 2	Example 3
n_1	317	231	131
n_2	284	238	69
median$_1$	7.34225	104	508
median$_2$	6.80445	100.5	565
$O\hat{V}L$	0.87714	0.85528	0.79234
$V\hat{a}r(O\hat{V}L)$ B=100	0.000919	0.000874	0.0018893
$V\hat{a}r(O\hat{V}L)$ B=200	0.000878	0.000929	0.0015787
$V\hat{a}r(O\hat{V}L)$ B=500	0.000823	0.000823	0.0016212
Lower 95% CL	0.816950	0.788187	0.704385
Upper 95% CL	0.901633	0.891234	0.846907

Note:

Example 1: Comparison of the median wealth for persistent and non-persistent Alabama farmers between 1850 and 1860.

Example 2: Comparison of the median Drumcondra Verbal Reasoning Test Score for Irish School children by gender in 1976.

Example 3: Comparison of the median minutes from onset of ischemic chest pain to ECG by gender for patients who experienced chest pain for more than six hours.

GRADUATE SCHOOL
UNIVERSITY OF ALABAMA AT BIRMINGHAM
DISSERTATION APPROVAL FORM

Name of Candidate Traci E. Clemons

Major Subject Biostatistics

Title of Dissertation A Nonparametric Approach to Estimating the Overlapping Coefficient Using the Kernel Estimation Technique

Dissertation Committee:

Edwin Bradley Jr, Chair

Mary E. Hovinga

Pauline E. Jolly

Director of Graduate Program

Dean, UAB Graduate School Joan Lorden

Date 6/7/97