1984

Behavior And Properties Of The Overlapping Coefficient As A Measure Of Agreement Between Distributions (Association, Dissimilarity).

Henry Forrest Inman
University of Alabama at Birmingham

INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of "sectioning" the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

University Microfilms International
300 N. Zeeb Road
Ann Arbor, MI 48106

Inman, Henry Forrest

BEHAVIOR AND PROPERTIES OF THE OVERLAPPING COEFFICIENT AS A MEASURE OF AGREEMENT BETWEEN DISTRIBUTIONS

The University of Alabama in Birmingham PH.D. 1984

PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark _√_.

1. Glossy photographs or pages _____
2. Colored illustrations, paper or print _____
3. Photographs with dark background _____
4. Illustrations are poor copy _____
5. Pages with black marks, not original copy _____
6. Print shows through as there is text on both sides of page _____
7. Indistinct, broken or small print on several pages _√_
8. Print exceeds margin requirements _____
9. Tightly bound copy with print lost in spine _____
10. Computer printout pages with indistinct print _____
11. Page(s) _____ lacking when material received, and not available from school or author.
12. Page(s) _____ seem to be missing in numbering only as text follows.
13. Two pages numbered _____. Text follows.
14. Curling and wrinkled pages _____
15. Other _____

BEHAVIOR AND PROPERTIES OF THE OVERLAPPING COEFFICIENT AS A MEASURE OF AGREEMENT BETWEEN DISTRIBUTIONS

by

HENRY FORREST INMAN

A DISSERTATION

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in the Department of
Biostatistics and Biomathematics in The Graduate School,
University of Alabama in Birmingham

BIRMINGHAM, ALABAMA

1984

GRADUATE SCHOOL
UNIVERSITY OF ALABAMA IN BIRMINGHAM
DISSERTATION APPROVAL FORM

Name of Candidate HENRY FORREST INMAN

Major Subject Statistics

Title of Dissertation BEHAVIOR AND PROPERTIES OF THE OVERLAPPING COEFFICIENT AS A MEASURE OF AGREEMENT BETWEEN DISTRIBUTIONS

Dissertation Committee:

Edwin Bradley, Jr., Chairman
Alfred G. Bartholomew
David C. Noffal
Charles R. Katholi
Daniel C. Hurst
Malcolm S. Turner

Director of Graduate Program Daniel C. Hurst

Dean, UAB Graduate School Robert M. Berger

Date 8-24-84

ABSTRACT OF DISSERTATION
GRADUATE SCHOOL, UNIVERSITY OF ALABAMA IN BIRMINGHAM

Degree Ph.D. Major Subject Statistics

Name of Candidate HENRY FORREST INMAN

Title BEHAVIOR AND PROPERTIES OF THE OVERLAPPING COEFFICIENT AS A MEASURE OF AGREEMENT BETWEEN DISTRIBUTIONS

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

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

where $f_1(x)$ and $f_2(x)$ are the probability density functions for the 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 various estimators of OVL are investigated in three situations: maximum-likelihood estimation of OVL when sampling from two normal distributions; nonparametric estimation of OVL using spline density estimates constructed from samples from two unspecified distributions; and estimation of OVL when

the two populations of interest, or samples from them, are represented by the rows in a 2 X C contingency table.

Using Monte Carlo techniques, it is discovered that the sample estimators of OVL in each of these circumstances exhibit downward bias, that is, the true overlap is underestimated, and that this bias increases as the similarity of the distributions from which the samples are obtained increases. In the normal distribution and 2 X C table cases, approximations to the variance of the estimators of OVL are derived. The approximate sampling distribution of the estimator of OVL between two normal distributions with common variance can be related to the folded-normal distribution, and confidence intervals for OVL can be constructed from the sampling distribution of the Mahalanobis distance. Bootstrap estimators of the sampling variance of estimators of OVL in quadratic spline and 2 X C cases are shown to be reasonable, and bootstrap methods of constructing confidence intervals for OVL are illustrated. The behavior of the sample estimators of OVL in all three situations suggests that OVL can serve as a valuable check on the meaningfulness of differences detected between the two distributions of interest by other statistical techniques, but that OVL itself should not be used to test for the equality of the two distributions compared.

Abstract Approved by: Committee Chairman Edwin Bradley, Jr.

Program Director David C. Hurst

Date 8-24-84 Dean of Graduate School Robert M. Harper

ACKNOWLEDGEMENTS

Although many people helped make the research summarized in this dissertation an enjoyable and rewarding experience, the contributions of several individuals stand out, and I should like to acknowledge their help here. The advice, encouragement, and friendship of Edwin L. Bradley, Jr., which antedates the formal start of my work in the Department of Biostatistics and Biomathematics, proved invaluable. My understanding of and work with spline density estimation owes much to Charles R. Katholi. David C. Hurst has offered a number of useful comments on the completed dissertation. My research would have been much more difficult without the skilled assistance of Tinker B. Dunbar, who located innumerable journal articles through interlibrary loan. The quality of the final manuscript was greatly improved by Dorothy H. Bradley, who found several embarassing errors. The staff of the University Computer Center eased my computational chores on many occasions; I particularly wish to thank Joyce Iannuzzi, Sharon Matthews, and Howard Rohdy. The dissertation could not have been completed without the computer time furnished by Deans Blaine A. Brownell and Kenneth J. Roozen of the Graduate School and Alfred A. Bartolucci, chairman of the Department of Biostatistics and Biomathematics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One

INTRODUCTION

Suppose we are given two probability distributions with probability (density) functions $f_1(x;\theta_1)$ and $f_2(x;\theta_2)$. If both distributions are of some common form indexed by the values of the parameters θ_1 and θ_2, the two distributions must differ if $\theta_1 \neq \theta_2$. However, θ_1 and θ_2 may differ and yet be similar in magnitude, suggesting that $f_1(x;\theta_1)$ and $f_2(x;\theta_2)$, while not identical, are similar. On the other hand, two distributions which are not of the same parametric form, say $f_1(x;\theta_1)$ and $f_2(x;\lambda_2)$, cannot be identical, but for certain values of the parameters θ_1 and λ_2 they may in fact be quite similar. Once again, the issue is the degree to which two distributions, known to differ, are similar or dissimilar.

A more realistic setting for this problem appears when the question of the similarity of two distributions is addressed through random samples selected from each of the two unknown distributions or populations. Assuming common form, $f_1(x;\theta_1)$ and $f_2(x;\theta_2)$ can be shown to differ by the appropriate statistical test for the equality of the parameters θ_1 and θ_2, given that $\theta_1 \neq \theta_2$. Since the power of such tests is usually related to both the magnitude of the difference in θ_1 and θ_2 and the sizes of the two samples from which θ_1 and θ_2 are estimated, small differences in θ_1 and θ_2 can be declared statistically

significant given sufficiently large sample sizes. Nevertheless, it is the magnitude of the estimated difference between θ_1 and θ_2, not the sizes of the samples, which actually indicates the degree of separation between $f_1(x;\theta_1)$ and $f_2(x;\theta_2)$. The prospect of declaring a trivial difference between θ_1 and θ_2 statistically significant while missing the true similarity of the two populations of interest as sample sizes increase has not been ignored. Commentators on statistical practice in many diverse areas of application have urged that the distinction between the statistical significance and the practical significance of differences detected in the parameters of the distributions that the populations of interest are presumed to follow be recognized (Boring, 1919; Cohen, 1962, 1977; Sheehan, 1980, for example), and introductory statistics textbooks often include a short discussion of the problem (for instance, Wallis and Roberts, 1956, pp. 384-85, 408-9; Snedecor and Cochran, 1980, p. 67; Moore, 1979, p. 292).

This study examines a measure of the agreement between two distributions proposed by Bradley and Piantadosi (1982) as a method of gauging the meaningfulness of some specified or estimated difference between the two probability distributions. This measure of agreement, the overlapping coefficient or OVL, indicates the similarity between the distributions of interest by computing--or estimating--the common area below the two probability densities; see figure 1.1. The greater the common area, the more similar are the two distributions. Bradley and Piantadosi determine OVL for several cases involving known distributions, but they do not consider the sampling behavior of estimators of OVL.

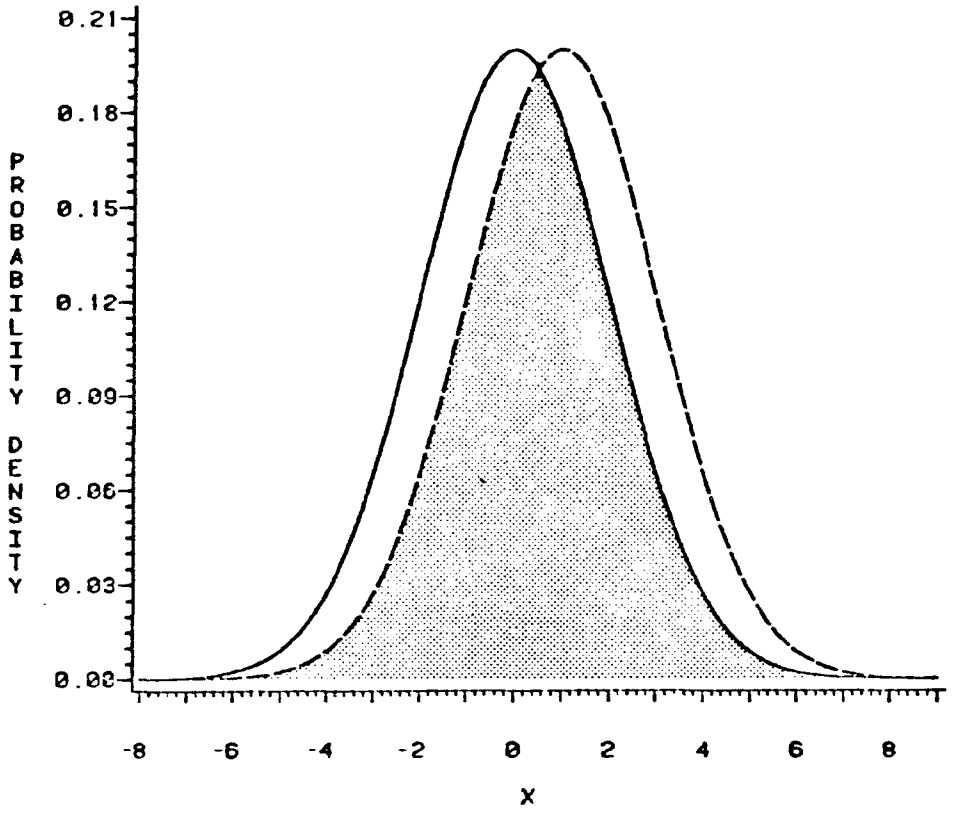

Figure 1.1 Graphical depiction of the overlapping coefficient. OVL is the shaded area in the figure.

Here the properties and sampling behavior of such estimators of OVL are investigated in three situations: when sampling from two normal distributions; when sampling from two distributions estimated non-parametrically by quadratic splines; and when samples from two discrete distributions are arranged in a 2 X C contingency table. It is discovered that the sample estimators of OVL in each of these circumstances are characterized by downward bias, that is, the true overlap is underestimated, and that this bias increases as the similarity between the distributions from which the samples are obtained increases. Further insight into the sampling behavior of the estimators of OVL is provided by Monte Carlo simulation studies in each of the three cases examined. In the normal distribution and the 2 X C table cases, estimates of the sample variance of the estimators of OVL can be derived, and normal approximations to the expectation and variance of the estimators of OVL in the 2 X C table are also presented. Bootstrap estimators of the sampling variance of the estimators of OVL in the quadratic spline and 2 X C table cases are shown to be reasonable, and bootstrap methods of constructing confidence intervals for OVL are illustrated. The behavior of the sample estimators of OVL in all three situations suggests that OVL can serve as a valuable check on the meaningfulness of differences detected between the two distributions of interest by other statistical techniques, but that OVL itself should not be used to test for the equality of the two distributions compared.

Definition of OVL

Let $f_1(x)$ and $f_2(x)$ be two probability (density) functions defined on some common domain for x. If $f_1(x)$ and $f_2(x)$ are continuous

distributions, then the overlapping coefficient is defined as

$$OVL = \int_x \min[f_1(x), f_2(x)]dx \ . \qquad (1.1)$$

If $f_1(x)$ and $f_2(x)$ are discrete distributions, then the overlapping coefficient is defined in an analagous manner:

$$OVL = \sum_x \min[f_1(x), f_2(x)] \ . \qquad (1.2)$$

As Bradley and Piantadosi indicate, OVL follows one of the usual conventions for measures of association (Goodman and Kruskal, 1979, p. 8). First, OVL always lies between zero and unity. Second, OVL attains unity if and only if the two distributions are identical. Finally, OVL is zero if and only if the two distributions being compared are totally distinct.

Relationship between OVL and the Index of Dissimilarity

OVL is directly related to a measure of association frequently used in the context of 2 X C contingency tables. The relationship between OVL and the index of dissimilarity, D, can be seen most easily if we rewrite the minimum of the two density functions, using the fact

that $f_1(x)$ and $f_2(x)$ are nonnegative:

$$\min[f_1(x), f_2(x)] = \frac{1}{2}\left[f_1(x) + f_2(x) - |f_1(x) - f_2(x)|\right] . \tag{1.3}$$

Substituting this expression into equations 1.1 and 1.2, we find

$$OVL = 1 - D ; \tag{1.4}$$

where D in the continuous case is given by

$$D = \frac{1}{2} \int_x |f_1(x) - f_2(x)| dx , \tag{1.5}$$

and in the discrete case by

$$D = \frac{1}{2} \sum_x |f_1(x) - f_2(x)| . \tag{1.6}$$

Thus the properties of OVL apply to D, except that D is zero when the two distributions compared are identical and unity when they are completely distinct. (D apparently always has been used in the discrete case and is usually defined as in equation 1.6.)

Calculation of OVL between Known Distributions

The method of determining the overlap between two specified distributions illustrates the general logic of computing OVL in any setting. Bradley and Piantadosi (1982) present as examples the overlap between two normal distributions, the overlap between the normal and the logistic distribution, and the overlap between two two-parameter exponential distributions. Here two additional examples are presented. In each, the computation of OVL is based on numerically or analytically determining $\min[f_1(x), f_2(x)]$.

OVL between the Standard Normal and Standard Cauchy Distributions

Here the overlapping coefficient between the standard normal distribution and the standard Cauchy distribution is computed.

The density of the standard normal random variable is

$$f_1(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \ ; \tag{1.7}$$

its distribution function is, of course, $\Phi(x)$.

The density of the standard Cauchy random variable is

$$f_2(x) = \left[\pi(1 + x^2)\right]^{-1} \; ; \tag{1.8}$$

its distribution function is

$$F_2(x) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1}(x) \; . \tag{1.9}$$

Since the two densities are symmetric about the point x = 0, we see that the two points of intersection of the densities are equidistant from zero; see figure 1.2. Thus we need only find one of these points, the lower crossing point, say, to evaluate OVL. If we equate the densities $f_1(x)$ and $f_2(x)$, we obtain the following nonlinear equation for the crossing points:

$$g(x) = \log_e(1 + x^2) - \frac{1}{2} x^2 + \frac{1}{2}\left[\log_e(\pi) - \log_e(2)\right] = 0 \; . \tag{1.10}$$

The derivative of this function with respect to x is

$$g'(x) = \frac{x(1 - x^2)}{1 + x^2} \; . \tag{1.11}$$

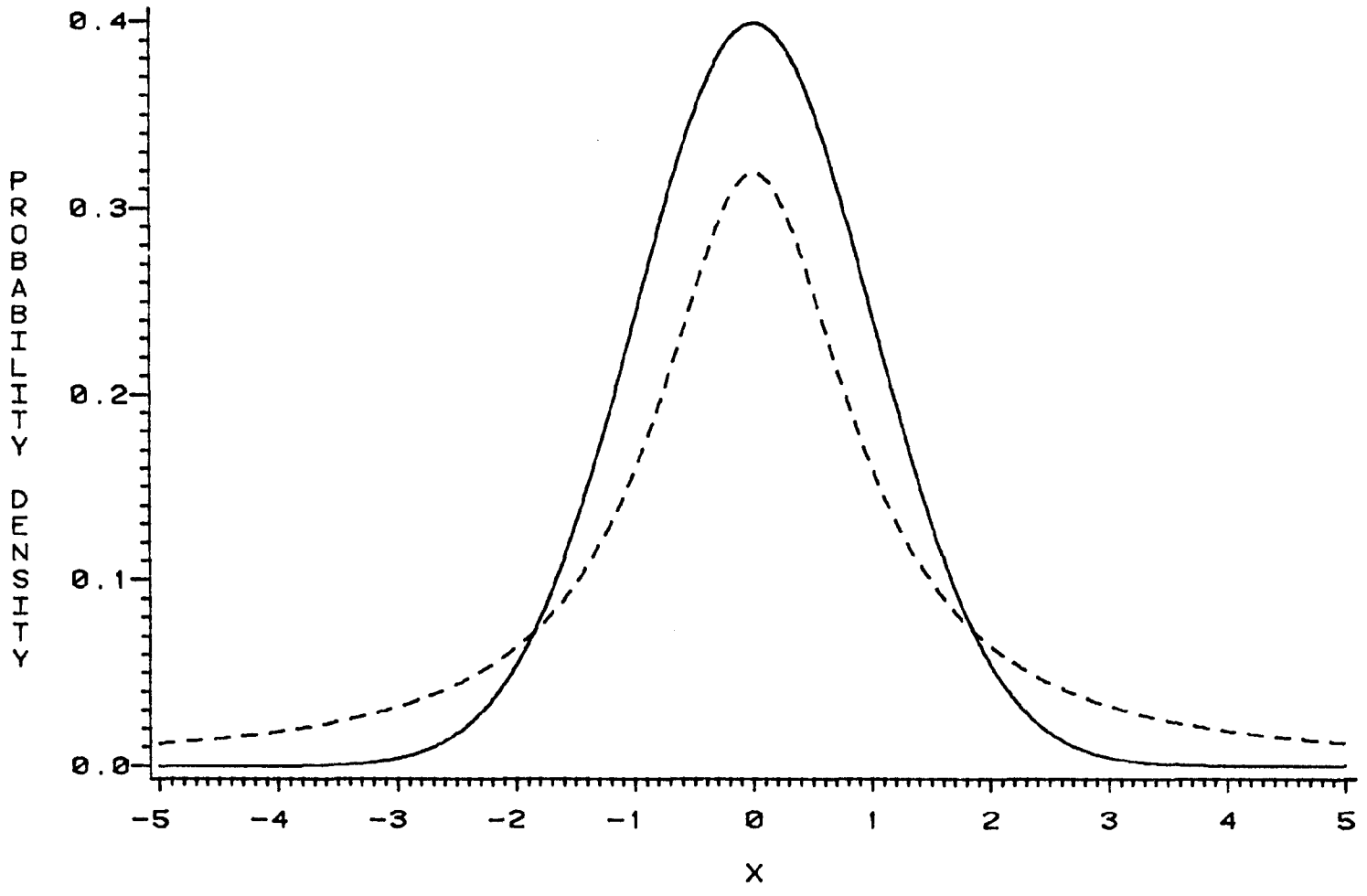

Figure 1.2 The standard normal and the standard Cauchy probability density functions. The standard normal density is indicated by the solid line, and the standard Cauchy density is shown by the broken line.

By Newton's method, we find that the two points at which the densities intersect are -1.851229 and 1.851229.

Using the symmetry of the two densities, we obtain for OVL:

$$OVL = 1 + 2\left[\Phi(-1.851229) - F_2(-1.851229)\right] = 0.748835 .$$

OVL between Two Poisson Distributions

Suppose we have two Poisson distributions with probability functions

$$P_i(x;\lambda_i) = \frac{\lambda_i^x \exp(-\lambda_i)}{x!}, \quad \lambda_i > 0;\ i=1,2;\ x=0,1,2\ldots . \tag{1.12}$$

By equating $P_1(x;\lambda_1)$ and $P_2(x;\lambda_2)$, we find a single "crossing point":

$$x_0 = \frac{\lambda_1 - \lambda_2}{\log_e(\lambda_1) - \log_e(\lambda_2)} . \tag{1.13}$$

Let $\lambda_1 > \lambda_2$. Then for $x \le x_0$, $P_1(x;\lambda_1) \le P_2(x;\lambda_2)$; for $x > x_0$, $P_1(x;\lambda_1) > P_2(x;\lambda_2)$.

Let $[x_0]$ denote the largest integer less than or equal to x_0. Then

$$OVL = \sum_{x=0}^{[x_0]} P_1(x;\lambda_1) + \sum_{x=[x_0]+1}^{\infty} P_2(x;\lambda_2)$$

$$= 1 - \sum_{x=0}^{[x_0]} \left[\frac{\lambda_2^x \cdot \exp(-\lambda_2) - \lambda_1^x \cdot \exp(-\lambda_1)}{x!} \right] . \qquad (1.14)$$

For example, if $\lambda_1 = 5$ and $\lambda_2 = 4$, then

$$x_0 = \frac{5 - 4}{\log_e(5) - \log_e(4)} = 4.4814 ,$$

and thus $[x_0] = 4$. Therefore OVL, computed from equation 1.14, is

$$OVL = 1 - \sum_{x=0}^{4} \left[\frac{4^x \cdot \exp(-4) - 5^x \cdot \exp(-5)}{x!} \right] = 0.811656 .$$

An Invariance Property of OVL

A useful property of OVL follows directly from equation 1.1. Let $g(x)$ be a continuous differentiable function defined for all x which is one-to-one and preserves order. Then OVL can be written in terms of $g(x)$ instead of x, based on integration with a change of variable, as

$$OVL = \int_{g(x)} \min[f_1(g(x)), f_2(g(x))]\, dg(x) \ . \qquad (1.15)$$

This invariance property of OVL is used explicitly in the development of the spline estimator of the overlapping coefficient in Chapter Three, but it is also allows immediate generalization of the results obtained for the estimation of OVL under normal theory to all cases where a normalizing transformation (Tukey, 1957; Box and Cox, 1964) can be found. An example of the latter instance is in fact presented in Chapter Two.

Previous Work Related to OVL

In its manifestation as D, the basic idea behind the overlapping coefficient extends back to the early years of the development of mathematical statistics. During the 1890s, Karl Pearson used a measure equivalent to 2D as an indicator of the goodness-of-fit of sample data to some theoretical distribution before his development of the technique based on the chisquare statistic (Pearson, 1965). Shortly after the second World War, the index of dissimilarity was reformulated several

times, apparently independently, by researchers in a number of disciplines (Duncan and Duncan, 1955). In the context of the 2 X C table, D is simply one of many proposed measures of association, and its general relationship to them is noted by Goodman and Kruskal (1979). More recently, interest in D appears to center on its use as an indicator of racial segregation and the probability model for the 2 X C table proposed by Cortese et al. (1976).

Weitzman (1970) seems to be the first analyst to work with OVL directly. He derived OVL in the discrete case from its relationship to the index of dissimilarity, and he used it to explore the differences in the income distributions of whites and blacks in the United States. Gastwirth (1973, 1975) briefly examined OVL and judged it inferior to a measure of the similarity of income distributions related to the Mann-Whitney form of the Wilcoxon test for equality of population medians. Gastwirth's objection to OVL is that it is unable to detect changes in the location of the common probability mass shared by the two distributions compared. Thus OVL was insensitive to shifts in the median income of women relative to that of men in the United States in his analysis of a longitudinal sample of Social Security records. Interest in OVL among statisticians in the United States appears to have ended with Gastwirth's critique.

Two investigators outside the United States have published recent material using the concept of the overlap of distributions in unrelated contexts. In Germany, the overlapping coefficient as a measure of association between two normal distributions with equal variances was developed by Marx (1976a, 1976b), and his proposal comes closest to the

form of OVL derived by Bradley and Piantadosi. Marx relies on the relationship of a sample estimator of the overlap between two identical normal distributions to the central t distribution (incorrectly specified) to produce a table of critical values for the sample overlapping coefficient. This, of course, accomplishes nothing, since Marx is simply transforming the scale of the usual t-test for the equality of the means of two normal populations. In addition, Marx assumes that because the sample realizations of OVL must lie between zero and unity, the sample overlapping coefficient can be treated as the usual sample estimator of a population proportion. Thus he compares two sample overlapping coefficients using the standard errors of sample estimators of population proportions and critical points from the central t distribution. Throughout, Marx averages sample sizes to obtain the degrees of freedom for the points of the t distributions he chooses to use. Unfortunately, then, there is nothing in Marx's work to increase our understanding of OVL, even in the simple case for which he proposes the use of the overlapping coefficient as a measure of association.

In Britain, Sneath (1977, 1979) has advanced the concept of overlap in the context of cluster analysis. Unlike Marx, Sneath correctly develops his treatment of the overlap of two normal distributions with equal population variances, but the correspondence between the overlap of two such normal distributions and the usual t-test for equality of normal population means apparently leads Sneath astray when he attempts to extend his results to the overlap between two normal distributions with unequal variances. While OVL has a direct interpretation in the problem of classifying individuals into two

populations, Sneath's clustering perspective does not speak directly to the nore general issue of comparing distributions which is addressed here.

Chapter Two

OVL BETWEEN TWO NORMAL POPULATIONS

The overlapping coefficient, OVL, between two normal distributions was derived by Bradley and Piantadosi (1982) for the equal and unequal population variances cases; they did not, however, discuss the estimation of OVL from sample data. Here the estimation of OVL using maximum-likelihood is addressed. The maximum-likelihood estimator of OVL, $\hat{OVL}$, is a biased estimator of OVL, and its bias depends directly on OVL itself: The bias of $\hat{OVL}$ increases as OVL nears one. As one should expect from the properties of maximum-likelihood estimators, the bias of $\hat{OVL}$ decreases as sample sizes become large, but this bias remains substantial when OVL is close to one even for large sample sizes. Estimates of the variance of $\hat{OVL}$, developed by the technique of statistical differentials, closely approximate the observed variance of $\hat{OVL}$ in a Monte Carlo experiment in two situations: when the population variances are equal and the difference in population means is small, and when the population variances are unequal and the difference in population means is large. The sampling distribution of the maximum-likelihood estimator can be related to the folded-normal distribution in the case of equal population variances and thus, for sufficiently large samples, to the normal distribution. Taken together, the properties of $\hat{OVL}$ observed in the Monte Carlo experiment provide realistic guidance to the actual

use of $\hat{OVL}$. In particular, the bias of $\hat{OVL}$ and the problem of estimating its variance accurately circumscribe the use of $\hat{OVL}$ as as an inferential statistic, suggesting that the proper role for $\hat{OVL}$ when sampling from normal distributions is similar to that of OVL when the two distributions are known. That is, $\hat{OVL}$ provides an indication of the meaningfulness of any difference in the normal distributions determined by the sample estimates of their means and variances, whatever the statistical significance of any differences in the estimated parameters.

The Overlap Between Two Known Normal Distributions

Suppose we are given two normal distributions with densities $f_1(x;\mu_1,\sigma_1^2)$ and $f_2(x;\mu_2,\sigma_2^2)$; that is, with means μ_1 and μ_2 and variances σ_1^2 and σ_1^2 respectively. From the general definition of the overlapping coefficient, one can determine OVL between these normal distributions in two cases of interest (Bradley and Piantadosi, 1982).

Equal Population Variances: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

In the case of equal population variances, the two normal densities intersect at a single point, x_0, half-way between the means μ_1 and μ_2, ignoring the coincidence of the densities at $-\infty$ and $+\infty$; see figure 2.1. That is,

$$x_0 = \frac{\mu_1 - \mu_2}{2} \quad . \tag{2.1}$$

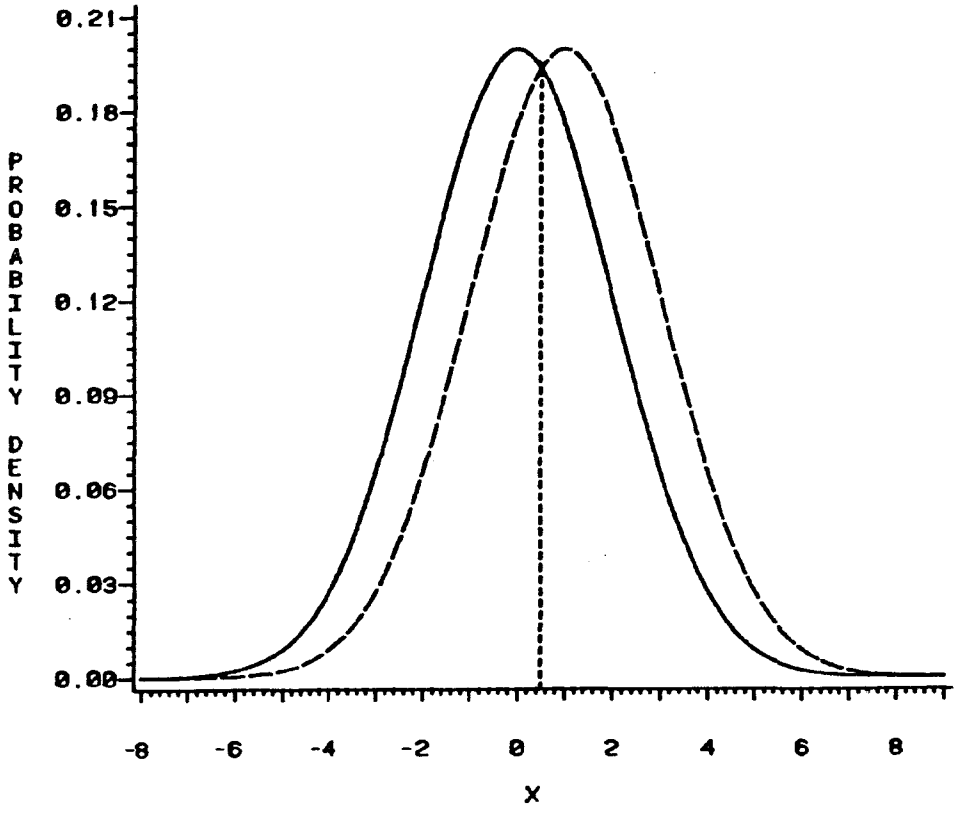

Figure 2.1 The overlap between two normal distributions with equal variances. The point of intersection, x_0, is indicated by the vertical broken line. Here $\mu_1 = 0$, $\mu_2 = 1$, $\sigma_1^2 = \sigma_1^2 = 4$, and $x_0 = 0.5$.

From the symmetry of $\min[f_1(x;\mu_1,\sigma_1^2),f_2(x;\mu_2,\sigma_2^2)]$ in this circumstance and the properties of the standard-normal distribution function, $\Phi(z)$, it is easy to see that

$$\text{OVL} = 2\Phi\left(\frac{-|\mu_1 - \mu_2|}{2\sigma}\right) \quad . \tag{2.2}$$

Thus, if $\mu_1 = 0$, $\mu_2 = 1$, and $\sigma^2 = 4$ (the situation depicted in figure 2.1), we compute $\text{OVL} = 2\Phi(-0.25) = 0.80258$.

Unequal Population Variances: $\sigma_1^2 \neq \sigma_2^2$

In the case of unequal population variances, the two normal densities--ignoring their coincidence at $-\infty$ and $+\infty$--intersect at exactly two points; see figure 2.2. These points are determined by the solutions to the quadratic equation in x obtained by setting the two densities equal to each other. If we assume $\sigma_2^2 > \sigma_1^2$, as in figure 2.2, the lower point of intersection, x_1, is given by

$$x_1 = \frac{\mu_1\sigma_2^2 - \mu_2\sigma_1^2 - \sigma_1\sigma_2\left[(\mu_1 - \mu_2)^2 + (\sigma_2^2 - \sigma_1^2)\log_e\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]^{1/2}}{\sigma_2^2 - \sigma_1^2} \quad , \tag{2.3}$$

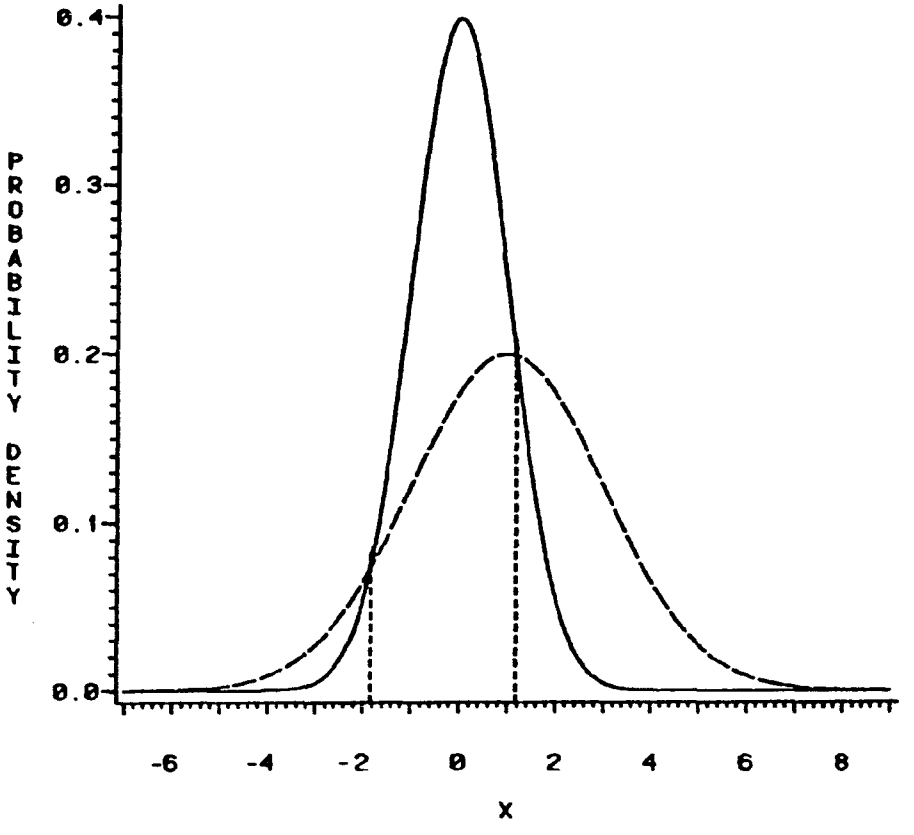

Figure 2.2 The overlap between two normal distributions with unequal variances. The lower point, x_1, and upper point, x_2, of intersection are indicated by the vertical broken lines. Here $\mu_1 = 0$, $\mu_2 = 1$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $x_1 = -1.847545$, and $x_2 = 1.180878$.

and the upper point of intersection, x_2, by

$$x_2 = \frac{\mu_1\sigma_2^2 - \mu_2\sigma_1^2 + \sigma_1\sigma_2\left[(\mu_1 - \mu_2)^2 + (\sigma_2^2 - \sigma_1^2)\log_e\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]^{\frac{1}{2}}}{\sigma_2^2 - \sigma_1^2} \quad . \tag{2.4}$$

If we define

$$z_{ij} = \frac{x_i - \mu_j}{\sigma_j} \quad , \; i=1,2; \; j=1,2; \tag{2.5}$$

then the overlap between the two distributions is given by the following equation:

$$OVL = \Phi(z_{11}) + \Phi(z_{22}) - \Phi(z_{12}) - \Phi(z_{21}) + 1 \quad . \tag{2.6}$$

Thus, for $\mu_1 = 0$, $\mu_2 = 1$, $\sigma_1^2 = 1$, and $\sigma_2^2 = 4$ (the situation illustrated

in figure 2.2), we can compute

$$x_1 = \frac{-1 - 2\sqrt{1 + \log_e(4)}}{3} = -1.847545,$$

and

$$x_2 = \frac{-1 + 2\sqrt{1 + \log_e(4)}}{3} = 1.180878.$$

Therefore

$$\begin{aligned} OVL &= \Phi(-1.847545) + \Phi(0.090439) - \Phi(-1.423773) - \Phi(1.180878) + 1 \\ &= 0.609934. \end{aligned}$$

Although it is not obvious, equation 2.6 reduces to equation 2.2 (in the limit) as $\sigma_2^2 \to \sigma_1^2$, or, equivalently, $\sigma_2 \to \sigma_1$. This becomes apparent when (2.3) and (2.4) are rewritten as the following.

$$x_1 = \frac{\sigma_2\mu_1 + \sigma_1\mu_2}{\sigma_2 + \sigma_1} + \frac{\sigma_1\sigma_2}{\sigma_2^2 - \sigma_1^2}\left\{\mu_1 - \mu_2 + \left[(\mu_1 - \mu_2)^2 + \right.\right.$$

$$\left.\left. + (\sigma_2^2 - \sigma_1^2)\log_e\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]^{1/2}\right\}, \tag{2.7}$$

$$x_2 = \frac{\sigma_2\mu_1 + \sigma_1\mu_2}{\sigma_2 + \sigma_1} + \frac{\sigma_1\sigma_2}{\sigma_2^2 - \sigma_1^2}\left\{\mu_1 - \mu_2 + \left[(\mu_1 - \mu_2)^2 + \right.\right.$$

$$\left.\left. + (\sigma_2^2 - \sigma_1^2)\log_e\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]^{1/2}\right\}. \tag{2.8}$$

Now because the product $(\sigma_2^2 - \sigma_1^2)\log_e\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$ converges to zero much faster than does $(\sigma_2^2 - \sigma_1^2)$ alone, we may write

$$\lim_{\sigma_2 \to \sigma_1}(x_1) = \frac{\mu_1 + \mu_2}{2} + \lim_{\sigma_2 \to \sigma_1}\left[\frac{\sigma_1\sigma_2}{\sigma_2^2 - \sigma_1^2}(\mu_1 - \mu_2 + |\mu_1 - \mu_2|)\right],$$

$$\lim_{\sigma_2 \to \sigma_1}(x_2) = \frac{\mu_1 + \mu_2}{2} + \lim_{\sigma_2 \to \sigma_1}\left[\frac{\sigma_1\sigma_2}{\sigma_2^2 - \sigma_1^2}(\mu_1 - \mu_2 - |\mu_1 - \mu_2|)\right].$$

Thus if $\mu_1 > \mu_2$,

$$\lim_{\sigma_2 \to \sigma_1}(x_1) = \frac{\mu_1 + \mu_2}{2} + \lim_{\sigma_2 \to \sigma_1}\left[\frac{2\sigma_1\sigma_2}{\sigma_2^2 - \sigma_1^2}(\mu_1 - \mu_2)\right] = +\infty ,$$

and

$$\lim_{\sigma_2 \to \sigma_1}(x_2) = \frac{\mu_1 + \mu_2}{2} + \lim_{\sigma_2 \to \sigma_1}\left[\frac{\sigma_1\sigma_2}{\sigma_2^2 - \sigma_1^2}(0)\right] = \frac{\mu_1 + \mu_2}{2} = x_0 .$$

Hence

$$\lim_{\sigma_2 \to \sigma_1 = \sigma}(OVL) = 1 + \Phi\left(\frac{x_0 - \mu_2}{\sigma}\right) - 1 - \Phi\left(\frac{x_0 - \mu_1}{\sigma}\right) + 1$$

$$= 2\Phi\left(\frac{-|\mu_1 - \mu_2|}{2\sigma}\right) .$$

On the other hand, if $\mu_1 < \mu_2$,

$$\lim_{\sigma_2 \to \sigma_1}(x_1) = \frac{\mu_1 + \mu_2}{2} + \lim_{\sigma_2 \to \sigma_1}\left[\frac{\sigma_1\sigma_2}{\sigma_2^2 - \sigma_1^2}\,(0)\right] = \frac{\mu_1 + \mu_2}{2} = x_0 \ .$$

and

$$\lim_{\sigma_2 \to \sigma_1}(x_2) = \frac{\mu_1 + \mu_2}{2} + \lim_{\sigma_2 \to \sigma_1}\left[\frac{2\sigma_1\sigma_2}{\sigma_2^2 - \sigma_1^2}\,(\mu_1 - \mu_2)\right] = -\infty \ .$$

Therefore

$$\lim_{\sigma_2 \to \sigma_1 = \sigma}(\mathrm{OVL}) = \Phi\left(\frac{x_0 - \mu_1}{\sigma}\right) + 0 - \Phi\left(\frac{x_0 - \mu_2}{\sigma}\right) - 0 + 1$$

$$= 2\Phi\left(\frac{-|\mu_1 - \mu_2|}{2\sigma}\right) \ .$$

Finally, let $\mu_1 = \mu_2 = \mu$. Obviously,

$$\lim_{\sigma_2 \to \sigma_1}(x_1) = \lim_{\sigma_2 \to \sigma_1}(x_2) = \frac{\mu_1 + \mu_2}{2} = \mu \ .$$

Hence

$$\lim_{\sigma_2 \to \sigma_1 = \sigma} (OVL) = \Phi(0) + \Phi(0) - \Phi(0) - \Phi(0) + 1 = 1 \ .$$

The convergence of OVL in the unequal variance case to OVL in the equal variance case is evident in table 2.1, which presents the value of OVL between two normal distributions for selected δ and γ, where

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma_1} \quad \text{and} \quad \gamma = \frac{\sigma_2^2}{\sigma_1^2} \ . \tag{2.9}$$

(Note that equations 2.2, 2.5, and 2.6 can all be written in terms of δ and γ instead of μ_1, μ_2, σ_1^2, and σ_2^2.)

Maximum-Likelihood Estimation of OVL

Now suppose that, instead of working with two known normal distributions, we have two independent simple random samples of sizes n_1 and n_2 from $f_1(x;\mu_1,\sigma_1^2)$ and $f_2(x;\mu_2,\sigma_2^2)$ respectively. Our problem is to estimate OVL from these sample data. Maximum-likelihood estimators of OVL can be derived simply in both the equal and the unequal variance cases by using the invariance property of the

TABLE 2.1

OVERLAPPING COEFFICIENT FOR TWO NORMAL DISTRIBUTIONS FOR SELECTED VALUES OF DELTA AND GAMMA

	GAMMA											
DELTA	1.00	1.10	1.20	1.30	1.40	1.50	1.75	2.00	2.25	2.50	2.75	3.00
0.00	1.00000	0.97694	0.95591	0.93661	0.91878	0.90222	0.86547	0.83394	0.80642	0.78208	0.76030	0.74064
0.05	0.98005	0.97116	0.95293	0.93461	0.91728	0.90103	0.86468	0.83335	0.80596	0.78169	0.75998	0.74036
0.10	0.96012	0.95667	0.94440	0.92875	0.91285	0.89748	0.86232	0.83159	0.80457	0.78055	0.75900	0.73952
0.15	0.94021	0.93874	0.93150	0.91942	0.90565	0.89166	0.85842	0.82868	0.80226	0.77864	0.75739	0.73812
0.20	0.92034	0.92002	0.91579	0.90722	0.89598	0.88373	0.85303	0.82464	0.79904	0.77599	0.75513	0.73617
0.25	0.90052	0.90109	0.89861	0.89288	0.88420	0.87389	0.84623	0.81951	0.79495	0.77259	0.75225	0.73367
0.30	0.88076	0.88208	0.88078	0.87710	0.87074	0.86242	0.83810	0.81332	0.78999	0.76848	0.74874	0.73062
0.35	0.86108	0.86307	0.86270	0.86045	0.85601	0.84957	0.82876	0.80613	0.78420	0.76366	0.74463	0.72705
0.40	0.84148	0.84408	0.84452	0.84333	0.84041	0.83564	0.81832	0.79800	0.77761	0.75816	0.73993	0.72296
0.45	0.82198	0.82515	0.82631	0.82597	0.82423	0.82089	0.80692	0.78900	0.77027	0.75201	0.73466	0.71836
0.50	0.80259	0.80630	0.80812	0.80852	0.80770	0.80555	0.79469	0.77920	0.76222	0.74522	0.72883	0.71327

TABLE 2.1 (CONTINUED)

	GAMMA											
DELTA	1.00	1.10	1.20	1.30	1.40	1.50	1.75	2.00	2.25	2.50	2.75	3.00
0.55	0.78332	0.78755	0.78998	0.79103	0.79098	0.78981	0.78176	0.76868	0.75350	0.73785	0.72247	0.70770
0.60	0.76418	0.76890	0.77190	0.77356	0.77418	0.77382	0.76827	0.75753	0.74417	0.72991	0.71560	0.70168
0.65	0.74518	0.75037	0.75391	0.75614	0.75736	0.75768	0.75434	0.74581	0.73428	0.72144	0.70825	0.69521
0.70	0.72634	0.73198	0.73602	0.73879	0.74056	0.74148	0.74006	0.73361	0.72388	0.71248	0.70044	0.68833
0.75	0.70766	0.71374	0.71826	0.72153	0.72381	0.72527	0.72553	0.72102	0.71303	0.70308	0.69221	0.68104
0.80	0.68916	0.69565	0.70063	0.70438	0.70714	0.70909	0.71084	0.70809	0.70178	0.69327	0.68358	0.67338
0.85	0.67084	0.67773	0.68314	0.68734	0.69056	0.69298	0.69604	0.69491	0.69020	0.68308	0.67458	0.66536
0.90	0.65271	0.65999	0.66581	0.67044	0.67409	0.67694	0.68120	0.68152	0.67832	0.67257	0.66524	0.65701
0.95	0.63479	0.64244	0.64865	0.65369	0.65775	0.66101	0.66635	0.66799	0.66619	0.66177	0.65560	0.64836
1.00	0.61708	0.62508	0.63166	0.63708	0.64154	0.64518	0.65153	0.65436	0.65388	0.65072	0.64568	0.63943
1.25	0.53197	0.54148	0.54966	0.55673	0.56285	0.56817	0.57867	0.58606	0.59075	0.59295	0.59302	0.59141
1.50	0.45325	0.46385	0.47320	0.48148	0.48885	0.49543	0.50908	0.51962	0.52773	0.53370	0.53770	0.53996

TABLE 2.1 (CONTINUED)

	GAMMA											
DELTA	1.00	1.10	1.20	1.30	1.40	1.50	1.75	2.00	2.25	2.50	2.75	3.00
1.75	0.38157	0.39284	0.40294	0.41204	0.42028	0.42776	0.44372	0.45658	0.46707	0.47562	0.48241	0.48760
2.00	0.31731	0.32883	0.33931	0.34887	0.35762	0.36567	0.38318	0.39770	0.40989	0.42021	0.42893	0.43621
2.25	0.26059	0.27200	0.28250	0.29218	0.30114	0.30945	0.32784	0.34340	0.35674	0.36828	0.37831	0.38703
2.50	0.21130	0.22229	0.23250	0.24202	0.25090	0.25922	0.27786	0.29393	0.30794	0.32024	0.33112	0.34078
2.75	0.16913	0.17945	0.18913	0.19823	0.20681	0.21490	0.23327	0.24938	0.26362	0.27631	0.28767	0.29790
3.00	0.13361	0.14307	0.15203	0.16053	0.16861	0.17630	0.19396	0.20970	0.22382	0.23654	0.24807	0.25856

maximum-likelihood estimators of the parameters μ_1, μ_2, σ_1^2, and σ_2^2 computed from the two samples. Maximum-likelihood theory insures that these estimators of OVL, $\hat{OVL}$, are asymptotically consistent, unbiased, efficient, and normally distributed (Kendall and Stuart, 1979, chap. 17). The approximate variances of $\hat{OVL}$ in the equal and unequal variance cases are derived using statistical differentials.

Equal Population Variances: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

The usual maximum-likelihood estimators for μ_1 and μ_2 are the sample means:

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}, \quad i=1,2; \; j=1,\ldots,n_i. \tag{2.10}$$

The variances of these estimators are given by

$$\mathrm{Var}(\bar{x}_i) = \frac{\sigma_i^2}{n_i}, \quad i=1,2. \tag{2.11}$$

The maximum-likelihood estimator of σ^2 can be written in the following way. Let

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n_i}, \quad i=1,2;\ j=1,\ldots,n_i. \tag{2.12}$$

Of course,

$$\mathrm{Var}(s_i^2) = \frac{2(n_i - 1)\sigma_i^4}{n_i^2}, \quad i=1,2. \tag{2.13}$$

Then the maximum-likelihood estimator of σ^2 $(= \sigma_1^2 = \sigma_2^2)$ is

$$s_p^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}; \tag{2.14}$$

its variance is given by

$$\mathrm{Var}(s_p^2) = \left(\frac{n_1}{n_1 + n_2}\right)^2 \mathrm{Var}(s_1^2) + \left(\frac{n_2}{n_1 + n_2}\right)^2 \mathrm{Var}(s_2^2)$$

$$= \frac{2(n_1 + n_2 - 2)\sigma^4}{(n_1 + n_2)^2} . \qquad (2.15)$$

Therefore, using equations 2.10 and 2.14, the maximum-likelihood estimator of OVL is the following:

$$\hat{OVL} = 2\Phi\left(\frac{-|\bar{x}_1 - \bar{x}_2|}{2s_p}\right) . \qquad (2.16)$$

To obtain the variance of $\hat{OVL}$, we note that $\bar{x}_1 - \bar{x}_2$ is normally distributed with mean $\mu_1 - \mu_2$ and variance $\frac{n_1 + n_2}{n_1 n_2}\sigma^2$. Thus the the random variable $|\bar{x}_1 - \bar{x}_2|$ has the folded-normal distribution with mean and variance (Leone et al., 1961):

$$E(|\bar{x}_1 - \bar{x}_2|) = \left[\frac{2(n_1 + n_2)}{n_1 n_2 \pi}\right]^{\frac{1}{2}} \cdot \sigma \cdot \exp\left[\frac{-n_1 n_2 (\mu_1 - \mu_2)^2}{2(n_1 + n_2)\sigma^2}\right] +$$

$$+ (\mu_1 - \mu_2)\left\{1 - 2\Phi\left[-\left(\frac{n_1 n_2}{n_1 + n_2}\right)^{\frac{1}{2}} \frac{\mu_1 - \mu_2}{\sigma}\right]\right\} , \qquad (2.17)$$

$$\mathrm{Var}(|\bar{x}_1 - \bar{x}_2|) = \frac{n_1 + n_2}{n_1 n_2}\sigma^2 + (\mu_1 - \mu_2)^2 - \left[\mathrm{E}(|\bar{x}_1 - \bar{x}_2|)\right]^2 . \qquad (2.18)$$

Then, by statistical differentials (Kendall and Stuart, 1977, pp. 246-47),

$$\mathrm{Var}(\hat{\mathrm{OVL}}) \doteq \left(\frac{\partial \hat{\mathrm{OVL}}}{\partial |\bar{x}_1 - \bar{x}_2|}\right)^2 \mathrm{Var}(|\bar{x}_1 - \bar{x}_2|) + \left(\frac{\partial \hat{\mathrm{OVL}}}{\partial s_p^2}\right)^2 \mathrm{Var}(s_p^2) , \qquad (2.19)$$

where the derivatives are understood to be evaluated at $\bar{x}_1 = \mu_1$, $\bar{x}_2 = \mu_2$, and $s_p^2 = \sigma^2$. But

$$\frac{\partial \hat{\mathrm{OVL}}}{\partial |\bar{x}_1 - \bar{x}_2|} = -\frac{1}{s_p}\phi\left(\frac{-|\bar{x}_1 - \bar{x}_2|}{2s_p}\right) , \qquad (2.20)$$

$$\frac{\partial \hat{\mathrm{OVL}}}{\partial s_p^2} = \frac{|\bar{x}_1 - \bar{x}_2|}{2 s_p^3}\phi\left(\frac{|\bar{x}_1 - \bar{x}_2|}{2s_p}\right) , \qquad (2.21)$$

where

$$\phi(z) = \frac{\exp(-z^2/2)}{\sqrt{2\pi}} . \tag{2.22}$$

Combining equations 2.15, 2.18, 2.19, 2.20, and 2.21, we obtain

$$\begin{aligned} \widehat{\mathrm{Var}(\mathrm{OVL})} &\doteq \phi^2\left(\frac{-|\mu_1 - \mu_2|}{2\sigma}\right)\left\{\frac{1}{\sigma^2}\,\mathrm{Var}(|\bar{x}_1 - \bar{x}_2|) + \frac{(\mu_1 - \mu_2)^2}{4\sigma^6}\,\mathrm{Var}(s_p^2)\right\} \\ &= \phi^2\left(\frac{-|\mu_1 - \mu_2|}{2\sigma}\right)\left\{\frac{n_1 + n_2}{n_1 n_2} + \right. \\ &\quad \left. + \left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2\left[1 + \frac{n_1 + n_2 - 2}{2(n_1 + n_2)}\right] - \left[\frac{E(|\bar{x}_1 - \bar{x}_2|)}{\sigma}\right]^2\right\} . \end{aligned} \tag{2.23}$$

In passing, we note that if $\mu_1 = \mu_2$, that is, if OVL = 1.0, equation 2.23 reduces to

$$\widehat{\mathrm{Var}(\mathrm{OVL})} \doteq \frac{1}{2}\left(\frac{n_1 + n_2}{n_1 n_2}\right)\frac{\pi - 2}{\pi} . \tag{2.24}$$

Ordinarily, $\text{Var}(\hat{OVL})$ must itself be estimated, substituting the sample estimates of the parameters μ_1, μ_2, and σ^2 into equation 2.17 to get $E(|\bar{x}_1 - \bar{x}_2|)$ and equation 2.23 to get $\hat{\text{Var}}(\hat{OVL})$. This gives the following computational formula.

$$\hat{\text{Var}}(\hat{OVL}) = \phi^2\left(\frac{-|\bar{x}_1 - \bar{x}_2|}{2s_p}\right)\left\{\frac{n_1 + n_2}{n_1 n_2} + \left(\frac{\bar{x}_1 - \bar{x}_2}{s_p}\right)^2\left[1 + \frac{n_1 + n_2 - 2}{2(n_1 + n_2)^2}\right] - \left[\frac{E(|\bar{x}_1 - \bar{x}_2|)}{s_p}\right]^2\right\} . \quad (2.25)$$

Unequal Population Variances: $\sigma_1^2 \neq \sigma_2^2$

In this case, the maximum-likelihood estimators $\bar{x}_1$, $\bar{x}_2$, s_1^2, and s_2^2 from equations 2.10 and 2.12 can be substituted for μ_1, μ_2, σ_1^2, and σ_2^2 in equations 2.3, 2.4, 2.5, and 2.6 to obtain the maximum-likelihood estimator of OVL. Assuming $s_2^2 > s_1^2$,

$$\hat{x}_1 = \frac{\bar{x}_1 s_2^2 - \bar{x}_2 s_1^2 - s_1 s_2\left[(\bar{x}_1 - \bar{x}_2)^2 + (s_2^2 - s_1^2)\log_e\left(\frac{s_2^2}{s_1^2}\right)\right]^{1/2}}{s_2^2 - s_1^2} \quad (2.26)$$

is the maximum-likelihood estimator of the lower point of intersection

of the two densities, and

$$\hat{x}_2 = \frac{\bar{x}_1 s_2^2 - \bar{x}_2 s_1^2 + s_1 s_2 \left[(\bar{x}_1 - \bar{x}_2)^2 + (s_2^2 - s_1^2) \log_e \left(\frac{s_2^2}{s_1^2} \right) \right]^{\frac{1}{2}}}{s_2^2 - s_1^2} \tag{2.27}$$

is the maximum-likelihood estimator of the upper point of intersection of the densities $f_1(x;\mu_1,\sigma_1^2)$ and $f_2(x;\mu_2,\sigma_2^2)$. The maximum-likelihood estimator of OVL, then, is

$$\hat{OVL} = \Phi(\hat{z}_{11}) + \Phi(\hat{z}_{22}) - \Phi(\hat{z}_{12}) - \Phi(\hat{z}_{21}) + 1 \ , \tag{2.28}$$

where

$$\hat{z}_{ij} = \frac{\hat{x}_i - \bar{x}_j}{s_j} \ , \ i=1,2; \ j=1,2. \tag{2.29}$$

In the case of unequal population variances, the technique of statistical differentials provides the following equation for the approximate variance of $\hat{OVL}$:

$$\mathrm{Var}(\hat{OVL}) \doteq \left(\frac{\partial \hat{OVL}}{\partial \bar{x}_1}\right)^2 \mathrm{Var}(\bar{x}_1) + \left(\frac{\partial \hat{OVL}}{\partial \bar{x}_2}\right)^2 \mathrm{Var}(\bar{x}_2) + \left(\frac{\partial \hat{OVL}}{\partial s_1^2}\right)^2 \mathrm{Var}(s_1^2) +$$

$$+ \left(\frac{\partial \hat{OVL}}{\partial s_2^2}\right)^2 \mathrm{Var}(s_2^2) \ ,$$

where the derivatives of $\hat{OVL}$ are understood to be evaluated at $\bar{x}_1 = \mu_1$, $\bar{x}_2 = \mu_2$, $s_1^2 = \sigma_1^2$, and $s_2^2 = \sigma_2^2$. It will be easier to write the expression above in this way:

$$\mathrm{Var}(\hat{OVL}) \doteq \left(\frac{\partial OVL}{\partial \mu_1}\right)^2 \mathrm{Var}(\bar{x}_1) + \left(\frac{\partial OVL}{\partial \mu_2}\right)^2 \mathrm{Var}(\bar{x}_2) + \left(\frac{\partial OVL}{\partial \sigma_1^2}\right)^2 \mathrm{Var}(s_1^2) +$$

$$+ \left(\frac{\partial OVL}{\partial \sigma_2^2}\right)^2 \mathrm{Var}(s_2^2) \ ;$$

that is, differentiating equation 2.6 with respect to the parameters instead of differentiating equation 2.28 with respect to $\bar{x}_1$, $\bar{x}_2$, s_1^2, and s_2^2 and then replacing $\bar{x}_1$, $\bar{x}_2$, s_1^2, and s_2^2 in the derivatives with μ_1, μ_2, σ_1^2, and σ_2^2. We have the following result.

$$\mathrm{Var}(\hat{OVL}) \doteq \left[\phi(z_{11})\frac{\partial z_{11}}{\partial \mu_1} + \phi(z_{22})\frac{\partial z_{22}}{\partial \mu_1} - \phi(z_{12})\frac{\partial z_{12}}{\partial \mu_1} - \right.$$

$$- \phi(z_{21})\frac{\partial z_{21}}{\partial \mu_1}\Big]^2 \mathrm{Var}(\bar{x}_1) + \Big[\phi(z_{11})\frac{\partial z_{11}}{\partial \mu_2} + \phi(z_{22})\frac{\partial z_{22}}{\partial \mu_2} -$$

$$- \phi(z_{12})\frac{\partial z_{12}}{\partial \mu_2} - \phi(z_{21})\frac{\partial z_{21}}{\partial \mu_2}\Big]^2 \mathrm{Var}(\bar{x}_2) + \Big[\phi(z_{11})\frac{\partial z_{11}}{\partial \sigma_1^2} +$$

$$+ \phi(z_{22})\frac{\partial z_{22}}{\partial \sigma_1^2} - \phi(z_{12})\frac{\partial z_{12}}{\partial \sigma_1^2} - \phi(z_{21})\frac{\partial z_{21}}{\partial \sigma_1^2}\Big]^2 \mathrm{Var}(s_1^2) +$$

$$+ \Big[\phi(z_{11})\frac{\partial z_{11}}{\partial \sigma_2^2} + \phi(z_{22})\frac{\partial z_{22}}{\partial \sigma_2^2} - \phi(z_{12})\frac{\partial z_{12}}{\partial \sigma_2^2} -$$

$$- \phi(z_{21})\frac{\partial z_{21}}{\partial \sigma_2^2}\Big]^2 \mathrm{Var}(s_2^2) \ . \qquad (2.30)$$

Here $\phi(z)$ is defined as in (2.22).

The derivatives of the z_{ij} (i=1,2; j=1,2) can be written most easily in terms of the derivatives of x_1 and x_2 with respect to the various parameters:

$$\left.\begin{aligned} \frac{\partial z_{11}}{\partial \mu_1} &= \frac{1}{\sigma_1}\left(\frac{\partial x_1}{\partial \mu_1} - 1\right), \quad \frac{\partial z_{12}}{\partial \mu_1} = \frac{1}{\sigma_2}\left(\frac{\partial x_1}{\partial \mu_1}\right), \\ \frac{\partial z_{21}}{\partial \mu_1} &= \frac{1}{\sigma_1}\left(\frac{\partial x_2}{\partial \mu_1} - 1\right), \quad \frac{\partial z_{22}}{\partial \mu_1} = \frac{1}{\sigma_2}\left(\frac{\partial x_2}{\partial \mu_1}\right); \end{aligned}\right\} \qquad (2.31)$$

$$\left.\begin{aligned} \frac{\partial z_{11}}{\partial \mu_2} &= \frac{1}{\sigma_1}\left(\frac{\partial x_1}{\partial \mu_2}\right), \quad \frac{\partial z_{12}}{\partial \mu_2} = \frac{1}{\sigma_2}\left(\frac{\partial x_1}{\partial \mu_2} - 1\right), \\ \frac{\partial z_{21}}{\partial \mu_2} &= \frac{1}{\sigma_1}\left(\frac{\partial x_2}{\partial \mu_2}\right), \quad \frac{\partial z_{22}}{\partial \mu_2} = \frac{1}{\sigma_2}\left(\frac{\partial x_2}{\partial \mu_2} - 1\right); \end{aligned}\right\} \quad (2.32)$$

$$\left.\begin{aligned} \frac{\partial z_{11}}{\partial \sigma_1^2} &= \frac{1}{\sigma_1}\left(\frac{\partial x_1}{\partial \sigma_1^2} - \frac{z_{11}}{2\sigma_1}\right), \quad \frac{\partial z_{12}}{\partial \sigma_1^2} = \frac{1}{\sigma_2}\left(\frac{\partial x_1}{\partial \sigma_1^2}\right), \\ \frac{\partial z_{21}}{\partial \sigma_1^2} &= \frac{1}{\sigma_1}\left(\frac{\partial x_2}{\partial \sigma_1^2} - \frac{z_{21}}{2\sigma_1}\right), \quad \frac{\partial z_{22}}{\partial \sigma_1^2} = \frac{1}{\sigma_2}\left(\frac{\partial x_2}{\partial \sigma_1^2}\right); \end{aligned}\right\} \quad (2.33)$$

and

$$\left.\begin{aligned} \frac{\partial z_{11}}{\partial \sigma_2^2} &= \frac{1}{\sigma_1}\left(\frac{\partial x_1}{\partial \sigma_2^2}\right), \quad \frac{\partial z_{12}}{\partial \sigma_2^2} = \frac{1}{\sigma_2}\left(\frac{\partial x_1}{\partial \sigma_2^2} - \frac{z_{12}}{2\sigma_2}\right), \\ \frac{\partial z_{21}}{\partial \sigma_2^2} &= \frac{1}{\sigma_1}\left(\frac{\partial x_2}{\partial \sigma_2^2}\right), \quad \frac{\partial z_{22}}{\partial \sigma_2^2} = \frac{1}{\sigma_2}\left(\frac{\partial x_2}{\partial \sigma_2^2} - \frac{z_{22}}{2\sigma_2}\right). \end{aligned}\right\} \quad (2.34)$$

Let $U(\mu_1,\mu_2,\sigma_1,\sigma_2) = \left[(\mu_1 - \mu_2)^2 + (\sigma_2^2 - \sigma_1^2)\log_e\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]^{\frac{1}{2}}$.

Then, from equations 2.3 and 2.4, we obtain the derivatives of x_1 and

x_2 with respect to μ_1, μ_2, σ_1^2, and σ_2^2. These derivatives are the following:

$$\frac{\partial x_1}{\partial \mu_1} = \frac{\sigma_2^2 - \sigma_1\sigma_2(\mu_1 - \mu_2)\left[U(\mu_1,\mu_2,\sigma_1,\sigma_2)\right]^{-1}}{\sigma_2^2 - \sigma_1^2}, \tag{2.35}$$

$$\frac{\partial x_2}{\partial \mu_1} = \frac{\sigma_2^2 + \sigma_1\sigma_2(\mu_1 - \mu_2)\left[U(\mu_1,\mu_2,\sigma_1,\sigma_2)\right]^{-1}}{\sigma_2^2 - \sigma_1^2}, \tag{2.36}$$

$$\frac{\partial x_1}{\partial \mu_2} = \frac{-\sigma_1^2 + \sigma_1\sigma_2(\mu_1 - \mu_2)\left[U(\mu_1,\mu_2,\sigma_1,\sigma_2)\right]^{-1}}{\sigma_2^2 - \sigma_1^2}, \tag{2.37}$$

$$\frac{\partial x_2}{\partial \mu_2} = \frac{-\sigma_1^2 - \sigma_1\sigma_2(\mu_1 - \mu_2)\left[U(\mu_1,\mu_2,\sigma_1,\sigma_2)\right]^{-1}}{\sigma_2^2 - \sigma_1^2}, \tag{2.38}$$

$$\frac{\partial x_1}{\partial \sigma_1^2} = \frac{1}{\sigma_2^2 - \sigma_1^2}\left\{-\mu_2 - \frac{\sigma_2}{2\sigma_1}U(\mu_1,\mu_2,\sigma_1,\sigma_2) + \right.$$
$$\left. + \frac{\sigma_1\sigma_2}{2}\left[\frac{\sigma_2^2 - \sigma_1^2}{\sigma_1^2} + \log_e\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]\left[U(\mu_1,\mu_2,\sigma_1,\sigma_2)\right]^{-1} + x_1\right\}, \tag{2.39}$$

$$\frac{\partial x_2}{\partial \sigma_1^2} = \frac{1}{\sigma_2^2 - \sigma_1^2}\left\{-\mu_2 + \frac{\sigma_2}{2\sigma_1}\, U(\mu_1,\mu_2,\sigma_1,\sigma_2) + \right.$$

$$\left. - \frac{\sigma_1\sigma_2}{2}\left[\frac{\sigma_2^2 - \sigma_1^2}{\sigma_1^2} + \log_e\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]\left[U(\mu_1,\mu_2,\sigma_1,\sigma_2)\right]^{-1} + x_2\right\}, \qquad (2.40)$$

$$\frac{\partial x_1}{\partial \sigma_2^2} = \frac{1}{\sigma_2^2 - \sigma_1^2}\left\{\mu_1 - \frac{\sigma_1}{2\sigma_2}\, U(\mu_1,\mu_2,\sigma_1,\sigma_2) - \right.$$

$$\left. - \frac{\sigma_1\sigma_2}{2}\left[\frac{\sigma_2^2 - \sigma_1^2}{\sigma_2^2} + \log_e\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]\left[U(\mu_1,\mu_2,\sigma_1,\sigma_2)\right]^{-1} - x_1\right\}, \qquad (2.41)$$

$$\frac{\partial x_2}{\partial \sigma_2^2} = \frac{1}{\sigma_2^2 - \sigma_1^2}\left\{\mu_1 + \frac{\sigma_1}{2\sigma_2}\, U(\mu_1,\mu_2,\sigma_1,\sigma_2) + \right.$$

$$\left. + \frac{\sigma_1\sigma_2}{2}\left[\frac{\sigma_2^2 - \sigma_1^2}{\sigma_2^2} + \log_e\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]\left[U(\mu_1,\mu_2,\sigma_1,\sigma_2)\right]^{-1} - x_2\right\}. \qquad (2.42)$$

Substituting these results into equation 2.30, we obtain the following formula for the approximate variance of $\hat{OVL}$ in the unequal population variance case.

$$\mathrm{Var}(\hat{OVL}) \doteq \left\{\left[\frac{\phi(z_{11})}{\sigma_1} - \frac{\phi(z_{12})}{\sigma_2}\right]\frac{\partial x_1}{\partial \mu_1} + \left[\frac{\phi(z_{22})}{\sigma_2} - \frac{\phi(z_{21})}{\sigma_1}\right]\frac{\partial x_2}{\partial \mu_1} + \right.$$

$$+\left\{\frac{\phi(z_{21})-\phi(z_{11})}{\sigma_1}\right\}^2\cdot\frac{\sigma_1^2}{n_1}+\left\{\left[\frac{\phi(z_{11})}{\sigma_1}-\frac{\phi(z_{12})}{\sigma_2}\right]\frac{\partial x_1}{\partial\mu_2}+\right.$$

$$+\left[\frac{\phi(z_{22})}{\sigma_2}-\frac{\phi(z_{21})}{\sigma_1}\right]\frac{\partial x_2}{\partial\mu_2}+\left.\frac{\phi(z_{12})-\phi(z_{22})}{\sigma_2}\right\}^2\cdot\frac{\sigma_2^2}{n_2}+$$

$$+\left\{\left[\frac{\phi(z_{11})}{\sigma_1}-\frac{\phi(z_{12})}{\sigma_2}\right]\frac{\partial x_1}{\partial\sigma_1^2}+\left[\frac{\phi(z_{22})}{\sigma_2}-\frac{\phi(z_{21})}{\sigma_1}\right]\frac{\partial x_2}{\partial\sigma_1^2}+\right.$$

$$+\left.\frac{\phi(z_{21})\cdot z_{21}-\phi(z_{11})\cdot z_{11}}{2\sigma_1^2}\right\}^2\cdot\frac{2(n_1-1)\sigma_1^4}{n_1^2}+$$

$$+\left\{\left[\frac{\phi(z_{11})}{\sigma_1}-\frac{\phi(z_{12})}{\sigma_2}\right]\frac{\partial x_1}{\partial\sigma_2^2}+\left[\frac{\phi(z_{22})}{\sigma_2}-\frac{\phi(z_{21})}{\sigma_1}\right]\frac{\partial x_2}{\partial\sigma_2^2}+\right.$$

$$+\left.\frac{\phi(z_{12})\cdot z_{12}-\phi(z_{22})\cdot z_{22}}{2\sigma_2^2}\right\}^2\cdot\frac{2(n_2-1)\sigma_2^4}{n_2^2}\,. \qquad (2.43)$$

Equation 2.43 gives the approximate variance of $\hat{OVL}$ when the parameters of the two normal distributions, μ_1, μ_2, σ_1^2, and σ_2^2, are known. In practice, of course, one would compute an estimate of this variance, $\hat{Var}(\hat{OVL})$, using the sample estimates $\bar{x}_1$, $\bar{x}_2$, s_1^2, and s_2^2 for μ_1, μ_2, σ_1^2, and σ_2^2 in this expression.

Monte Carlo Investigation of the Properties of $\hat{OVL}$

Because the sample estimators of OVL given in equation 2.14 (equal population variances) and equation 2.28 (unequal population variances) are the maximum-likelihood estimators of OVL, they have known asymptotic properties: consistency, unbiasedness, efficiency,

and normality. Maximum-likelihood theory, of course, does not guarantee that the maximum-likelihood estimators of OVL exhibit these properties when they are based on small samples nor that these properties are attained rapidly as sample sizes increase. Since the distribution of $\hat{OVL}$ in either of the two cases of interest is not immediately evident from (2.14) or (2.28), the distributional properties of $\hat{OVL}$ are not obvious. Here, the basic statistical properties of $\hat{OVL}$ shall be determined in a Monte Carlo simulation study. The three objectives of this study are to investigate the sampling distribution of $\hat{OVL}$ and, in particular, the bias of $\hat{OVL}$ as an estimator of OVL; to examine the usefulness of the approximation formulae in (2.23) and (2.43) for the variance of $\hat{OVL}$; and, if possible, to determine the form of the sampling distribution of $\hat{OVL}$.

The Monte Carlo study itself can be described very briefly. For convenience, the first normal distribution is fixed at the standard normal, that is, with mean $\mu_1 = 0.00$ and variance $\sigma_1^2 = 1.0$. The mean μ_2 and variamce σ_2^2 of the second normal distribution are then selected to create seven design points for the study. These points were chosen to permit investigation of the properties of $\hat{OVL}$ in the following circumstances: 1) Two normal distributions with the same means and variances ($\mu_2 = 0.00$, $\sigma_2^2 = 1.0$); 2) two normal distributions with the same variance but unequal means, where this difference is small and large ($\mu_2 = 0.25$ and 1.00, $\sigma_2^2 = 1.0$); 3) two normal distributions with identical means but unequal variances, where this difference is small and large ($\mu_2 = 0.00$, $\sigma_2^2 = 1.2$ and 3.0); and 4) two normal distributions with unequal means and unequal variances, where both differences are small and large ($\mu_2 = 0.25$, $\sigma_2^2 = 1.2$ and $\mu_2 = 1.00$, $\sigma_2^2 = 3.0$). At each

of these seven design points, the sampling distribution of $\hat{OVL}$ is simulated for four sets of sample sizes for the independent samples from each distribution: $n_1 = n_2 = 50$, 100, 250, and 500. One thousand Monte Carlo trials were run for each set of sample sizes at each design point in the study, as follows. In each trial, two samples of standard-normal random deviates of the required size were generated using the IMSL routine GGNML (IMSL, 1982), employing different seeds for the two samples. The second set of standard-normal random deviates was then transformed to the desired mean μ_2 and variance σ_2^2. Then the sample means and sample variance estimates were calculated for each of the two samples, using the West algorithm (Chan and Lewis, 1979, p. 528). The sample overlapping coefficient, $\hat{OVL}$, was then calculated, using these estimates of the population means and variances, from equation 2.14 if $\sigma_2^2 = 1.0$ and from equation 2.28 if $\sigma_2^2 \neq 1.0$. Thus at each design point and for four sets of sample sizes we have 1000 Monte Carlo observations of $\hat{OVL}$. (All computer routines used in the Monte Carlo study can be found in the appendix.)

Bias and Predicted Variance of $\hat{OVL}$

The results of the Monte Carlo simulation experiment are presented in table 2.2. The true overlap between the two normal distributions, OVL, is calculated from equation 2.2 or equation 2.6 as appropriate, using the assigned values of μ_1, μ_2, σ_1^2, and σ_2^2. The predicted variance shown in the table is computed from equation 2.23 (if $\sigma_2^2 = 1.0$) or equation 2.43 (if $\sigma_2^2 \neq 1.0$), also using the assigned values of μ_1, μ_2, σ_1^2, and σ_2^2. The Monte Carlo mean and variance are computed from the observed first and second sample moments from the

TABLE 2.2

RESULTS OF MONTE CARLO SIMULATION STUDY: MAXIMUM LIKELIHOOD ESTIMATOR OF OVL BASED ON INDEPENDENT SAMPLES FROM TWO NORMAL DISTRIBUTIONS

$N_1=N_2$	PREDICTED VARIANCE	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED BIAS	VARIANCE RATIO
		$\sigma_2^2=1.0$, $\mu_2=0.00$, OVL=1.000000			
50	0.00231335	0.934009	0.00224669	-1.39224	0.971184
100	0.00115668	0.954478	0.00111680	-1.36217	0.965524
250	0.00046267	0.973020	0.00042176	-1.31375	0.911569
500	0.00023134	0.980415	0.00023081	-1.28916	0.997714
		$\sigma_2^2=1.0$, $\mu_2=0.25$, OVL=0.900524			
50	0.00466606	0.891624	0.00502834	-0.12551	1.077640
100	0.00281206	0.900347	0.00282416	-0.00333	1.004301
250	0.00125241	0.899449	0.00133316	-0.02942	1.064482
500	0.00063155	0.900285	0.00064187	-0.00941	1.016342

TABLE 2.2 (CONTINUED)

$N_1=N_2$	PREDICTED VARIANCE	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED BIAS	VARIANCE RATIO
		$\sigma_2^2=1.0$, $\mu_2=1.00$, OVL=0.617075			
50	0.00556535	0.611304	0.00547252	-0.07801	0.983321
100	0.00278578	0.615567	0.00284879	-0.02825	1.022620
250	0.00111505	0.617175	0.00115729	0.00293	1.037881
500	0.00055765	0.617527	0.00056256	0.01905	1.008801
		$\sigma_2^2=1.2$, $\mu_2=0.00$, OVL=0.955913			
50	0.00457128	0.899412	0.00260552	-1.10690	0.569976
100	0.00230896	0.922447	0.00145593	-0.87707	0.630557
250	0.00092918	0.942397	0.00071949	-0.50386	0.774328
500	0.00046552	0.949600	0.00035333	-0.33582	0.758983
		$\sigma_2^2=1.2$, $\mu_2=0.25$, OVL=0.898605			
50	0.00586572	0.868436	0.00380814	-0.48889	0.649218
100	0.00293499	0.888259	0.00256723	-0.20421	0.874699
250	0.00117451	0.893001	0.00100947	-0.17638	0.859489
500	0.00058734	0.896736	0.00054757	-0.07987	0.932284

TABLE 2.2 (CONTINUED)

$N_1 = N_2$	PREDICTED VARIANCE	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED BIAS	VARIANCE RATIO
		$\sigma_2^2=3.0$, $\mu_2=0.00$, OVL=0.740639			
50	0.00395722	0.731205	0.00383740	-0.15229	0.969721
100	0.00199880	0.734808	0.00187022	-0.13483	0.935671
250	0.00080437	0.739085	0.00087185	-0.05263	1.083904
500	0.00040299	0.739620	0.00039746	-0.05111	0.986285
		$\sigma_2^2=3.0$, $\mu_2=1.00$, OVL=0.639429			
50	0.00402558	0.628437	0.00395324	-0.17482	0.982029
100	0.00201991	0.633837	0.00205962	-0.12322	1.019656
250	0.00080967	0.636917	0.00078805	-0.08948	0.973286
500	0.00040512	0.638793	0.00044562	-0.03012	1.099972

1000 simulated $\hat{OVL}$ in each design-point-sample-size combination. Comparisons of the Monte Carlo mean to the true OVL indicate the bias of $\hat{OVL}$ as an estimator of OVL. Comparing the Monte Carlo variance to the variance predicted from the two approximation formulae demonstrates the utility of these equations in the best possible circumstance, when the parameters of the two distributions sampled are known. These comparisons are made explicitly in table 2.2 through the calculation of the standardized bias and the variance ratio. The standardized bias is simply the difference, Monte Carlo mean $\hat{OVL}$ minus OVL, divided by the square-root of the Monte Carlo variance. The variance ratio is the ratio of the Monte Carlo variance to the predicted variance.

The Monte Carlo experiment clearly demonstrated that OVL is biased: In general, $\hat{OVL}$ will understate OVL for the values of OVL considered here. As we should expect, this bias decreases as sample sizes increase, but this decrease in bias is apparently not monotone (see $\mu_2 = 0.25$, $\sigma_2^2 = 1.0$, $n_1 = n_2 = 100$ in table 2.2). The bias of $\hat{OVL}$ is also directly related to the value of OVL. The largest bias observed in the simulation study occurs when OVL = 1.0, and the bias of $\hat{OVL}$ decreases the further OVL is from unity. Evidently, the more similar the two normal distributions from which the two samples are drawn, the greater is the bias of $\hat{OVL}$ as an estimator of OVL.

The usefulness of the approximation formulae for the variance of OVL also appears to be related to the value of OVL. From the ratio of the Monte Carlo variance to the predicted variance, we see that in the equal population variance case, equation 2.23 performs well. In every set of Monte Carlo trials where $\sigma_2^2 = 1.0$, the ratio of the two variances is very near unity. On the other hand, in the unequal population

case, equation 2.43 performs best when OVL is distant from one and breaks down when $\mu_1 = \mu_2$. Moreover, it is evident from the Monte Carlo simulation study that the expression for the approximate variance of $\hat{OVL}$ in the unequal population variances case overstates the apparent sampling variance of $\hat{OVL}$ when it fails.

The Sampling Distribution of $\hat{OVL}$

The properties of $\hat{OVL}$ observed in the Monte Carlo study lead directly to the identification of the approximate distribution of $\hat{OVL}$, at least in the equal variance case. In fact, the Monte Carlo results suggest that, when the population variances are equal, the sampling distribution of $\hat{OVL}$ can be simply related to the folded-normal distribution. The justification for such a link can be made as follows. First, suppose that σ is known; $\hat{OVL}$ then becomes a function of the absolute difference in sample means only. This absolute difference, as we have already seen, follows the folded-normal distribution. If $\Phi(z)$ is viewed as an approximately linear transformation of this absolute difference, then $\hat{OVL}$ must also be related to this distribution. Naturally, as sample sizes increase and s_p provides a better estimate of σ, $\hat{OVL}$ should behave increasingly like this idealization. Thus as sample sizes increase and the sampling variances of s_p and $\hat{OVL}$ decrease, the assumptions about σ and $\Phi(z)$ become more reasonable, and we should then expect that $\hat{OVL}$ can be linearly related to some folded-normal random variable. Second, from the Monte Carlo simulation is appears that the bias of $\hat{OVL}$ diminishes with the distance of OVL, in units of the standard error of $\hat{OVL}$, from one, and that $\hat{OVL}$ exhibits a normal sampling distribution when OVL is sufficiently far from one. Of course,

$\hat{OVL}$, like OVL, is bounded above by 1.0, and in the Monte Carlo study $\hat{OVL}$ tends to "bunch" below 1.0 when OVL is near unity. This bunching is most severe when OVL = 1.0, and this behavior seems to account for the observed bias of $\hat{OVL}$. This suggests that the distribution of $\hat{OVL}$ is folded about the point 1.0. Using an obvious notation based on the relationship of OVL to the index of dissimilarity, let D = 1 - OVL and $\hat{D} = 1 - \hat{OVL}$. The statistic $\hat{D}$, in fact, follows the folded-normal distribution in the case of equal population variances.

The folded-normal distribution

The folded-normal distribution arises in the following way. Let the random variable x be normally distributed with mean ξ and variance τ^2. The random variable $y = |x|$ has the folded-normal distribution, a fact used to derive equation 2.23 above. The distribution of y is completely specified if ξ and τ are known, and the first and second noncentral moments of y are

$$\mu_1' = \mu_f = \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \cdot \tau \cdot \exp\left(\frac{-\xi^2}{2\tau^2}\right) + \xi\left[1 - 2\Phi\left(\frac{-\xi}{\tau}\right)\right] , \qquad (2.44)$$

and

$$\mu_2' = \tau^2 + \xi^2 . \qquad (2.45)$$

Thus the variance of y is $\sigma_f^2 = \tau^2 + \xi^2 - \mu_f^2$. (The subscript f is used to identify the mean and variance of the folded-normal variate.) In the special case when $\xi = 0$ (the half-normal distribution), we note that $\mu_f = \left(\frac{2}{\pi}\right)^{\frac{1}{2}}\tau$ and $\sigma_f^2 = \tau^2\left(1 - \frac{2}{\pi}\right)$. Clearly, then, $\mu_f \geq \left(\frac{2}{\pi}\right)^{\frac{1}{2}}\tau$ and $\tau^2\left(1 - \frac{2}{\pi}\right) \leq \sigma_f^2 \leq \tau^2$ must hold for any folded-normal variable. Elandt (1961, p. 554) notes that the folded-normal random variable converges to the normal distribution as ξ/τ increases, achieving approximate normality when $\mu_f/\sigma_f > 3$. Properties of the folded-normal distribution, including higher moments and its tabulated distribution, are discussed in Leone et al. (1961) and Elandt (1961). The folded-normal distribution is directly linked to the noncentral chisquare distribution. Let $\lambda = \xi^2/\tau^2$. Then the random variable y/τ is distributed as $\chi(1,\lambda)$, and the random variable y^2/τ^2 is distributed as $\chi^2(1,\lambda)$; see Krishnaiah et al. (1963) and Johnson and Kotz (1970, p. 136).

Estimation of ξ and τ is considered in Elandt (1961), Johnson (1962), and Johnson and Kotz (1970, pp. 136-37). Here ξ and τ will be estimated from the first and second Monte Carlo moments of $\hat{D}$, m_1' and m_2' respectively. Specifically, m_2' is equated to μ_2' in equation 2.45, giving

$$\xi = (m_2' - \tau^2)^{\frac{1}{2}} . \qquad (2.46)$$

This expression is substituted into equation 2.44, yielding the following nonlinear equation in τ:

$$f(\tau) = \mu_f - m_1' = 0 \ . \tag{2.47}$$

Then

$$f'(\tau) = \left(\frac{2}{\pi}\right)^{1/2}\left(1 + \frac{m_2'}{\tau^2}\right)\exp\left(\frac{-\xi^2}{2\tau^2}\right) - \frac{\tau}{\xi}\left[1 - 2\Phi\left(\frac{-\xi}{\tau}\right)\right] - \frac{2m_2'}{\tau^2}\,\phi\left(\frac{-\xi}{\tau}\right) \ , \tag{2.48}$$

where ξ is defined as in (2.46). Newton's method then permits estimation of τ, and thus ξ, by finding the solution to equation 2.47 subject to the condition that $\xi \geq 0$. (Since ξ in the denominator of the second term in equation 2.48 proves awkward when ξ is close to or equal to zero, the successive terms of Newton's method are more stably determined by

$$\tau_{n+1} = \tau_n - \frac{\xi \cdot f(\tau)}{\xi \cdot f'(\tau)} \ ,$$

which of course is algebraically equivalent to the usual formulation.) A convenient initial value for the estimation of τ is σ_f.

Goodness of fit of $\hat{D}$ to the folded-normal distribution

Estimates of ξ and τ are presented in table 2.3 for each set of trials in the Monte Carlo simulation study. To determine whether the folded-normal distribution adequately characterizes the sampling distribution of D, λ is calculated from ξ and τ, and the Kolmogorov statistic is used to test the equivalence of the empirical distribution function for $\hat{D}^2/\tau^2$ to the noncentral chisquare distribution function with a single degree of freedom and noncentrality parameter λ. The theoretical cumulative probabilities were calculated using the CPROB function in SAS (Hardison et al., 1983) or, when this proved unstable, the MDCHN routine in IMSL (1982). The Kolmogorov statistic, computed in the usual way (Gibbons, 1971, pp. 75-85), is then compared to the table of pseudocritical values in Stephans (1974) for testing normality when mean and variance are unknown. These statistics, together with similar Kolmogorov test statistics for normality, are presented in the last two columns of table 2.3. Obviously, there is no reason to reject the hypothesis that $\hat{D}$ follows the folded-normal distribution when $\sigma_1^2 = \sigma_2^2 = 1.0$. However, the folded-normal fails to represent the observed (Monte Carlo) sampling distribution of $\hat{D}$ when $\sigma_2^2 \neq 1.0$, agreeing only when $\hat{D}$ appears to be normally distributed. The standardized bias in table 2.3 is defined as $\frac{\xi - D}{\tau}$, and it indicates that ξ approaches D as n_1 and n_2 increase. Therefore, the Monte Carlo

TABLE 2.3

RESULTS OF MONTE CARLO SIMULATION STUDY: FITS OF MAXIMUM LIKELIHOOD ESTIMATOR OF OVL
BASED ON INDEPENDENT SAMPLES FROM TWO NORMAL DISTRIBUTIONS TO THE FOLDED-NORMAL AND NORMAL DISTRIBUTIONS

$N_1=N_2$	μ_f	σ_f	ξ	τ	λ	STANDARDIZED BIAS	KOLMOGOROV STATISTICS $\chi^2(1,\lambda)$	$N(\mu_f,\sigma_f^2)$
				$\sigma_2^2=1.0$, $\mu_2=0.00$, D=0.000000				
50	0.065991	0.0473992	.0505924	0.0635762	0.63326	0.79578	0.015602	0.081849***
100	0.045522	0.0334185	.0312611	0.0470294	0.441844	0.66471	0.010044	0.085881***
250	0.027054	0.0205367	1.8E-05	0.0339069	2.8E-07	0.00053	0.017282	0.096283***
500	0.019777	0.0151923	1.6E-05	0.0247869	4.1E-07	0.00064	0.025244	0.099181***
				$\sigma_2^2=1.0$, $\mu_2=0.25$, D=0.099476				
50	0.108376	0.0709108	.0980511	0.0846152	1.34279	-0.01685	0.018779	0.062659***
100	0.099653	0.0531428	.0977028	0.0566487	2.97463	-0.03131	0.018960	0.035208***
250	0.100551	0.0365125	0.100482	0.0367012	7.49577	0.02740	0.021710	0.023426
500	0.099715	0.0253352	.0997144	0.0253370	15.4883	0.00939	0.018039	0.018019

TABLE 2.3 (CONTINUED)

$N_1=N_2$	μ_f	σ_f	ξ	τ	λ	STANDARDIZED BIAS	KOLMOGOROV STATISTICS $\chi^2(1,\lambda)$	$N(\mu_f,\sigma_f^2)$
			$\sigma_2^2=1.0$, $\mu_2=1.00$, D=0.382925					
50	0.388696	0.0739765	0.388696	0.0739765	27.6079	0.07801	0.020691	0.020691
100	0.384433	0.0533741	0.384433	0.0533741	51.8777	0.02825	0.016679	0.016679
250	0.382825	0.0340190	0.382825	0.0340190	126.636	-0.00293	0.017118	0.017118
500	0.382473	0.0237183	0.382473	0.0237183	260.036	-0.01905	0.020257	0.020370
			$\sigma_2^2=1.2$, $\mu_2=0.00$, D=0.044087					
50	0.100588	0.0510443	.0992442	0.0536110	3.42692	1.02883	0.036006***	0.051535***
100	0.077553	0.0381567	.0767397	0.0397677	3.72374	0.82108	0.033432***	0.046418***
250	0.057603	0.0268234	.0572142	0.0276425	4.28403	0.47488	0.029013**	0.038165***
500	0.050400	0.0187969	.0503544	0.0189183	7.0845	0.33127	0.016631	0.018180
			$\sigma_2^2=1.2$, $\mu_2=0.25$, D=0.101395					
50	0.131564	0.0617101	0.130623	0.0636788	4.20773	0.45899	0.033132***	0.040901***
100	0.111741	0.0506679	0.111136	0.0519818	4.57098	0.18740	0.026718*	0.027398*
250	0.106999	0.0317722	0.106992	0.0317935	11.3248	0.17606	0.020404	0.020394
500	0.103264	0.0234001	0.103263	0.0234003	19.4737	0.07986	0.016062	0.016064

TABLE 2.3 (CONTINUED)

$N_1=N_2$	μ_f	σ_f	ξ	τ	λ	STANDARDIZED BIAS	KOLMOGOROV STATISTICS $\chi^2(1,\lambda)$	$N(\mu_f,\sigma_f^2)$
			$\sigma_2^2=3.0$, $\mu_2=0.00$, D=0.259361					
50	0.268795	0.0619467	0.268794	0.0619475	18.8275	0.15228	0.019406	0.019410
100	0.265192	0.0432460	0.265192	0.0432460	37.6034	0.13483	0.019014	0.019014
250	0.260915	0.0295272	0.260915	0.0295272	78.0824	0.05263	0.015883	0.015883
500	0.260380	0.0199365	0.26038	0.0199365	170.576	0.05111	0.018459	0.018569
			$\sigma_2^2=3.0$, $\mu_2=1.00$, D=0.360571					
50	0.371563	0.0628748	0.371563	0.0628748	34.923	0.17482	0.020517	0.020517
100	0.366163	0.0453830	0.366163	0.0453830	65.0972	0.12322	0.021973	0.021973
250	0.363083	0.0280721	0.363083	0.0280721	167.286	0.08948	0.017641	0.017789
500	0.361207	0.0211098	0.361207	0.0211098	292.782	0.03012	0.024334	0.024494

NOTE: ASTERISKS DENOTE REJECTION OF THE INDICATED DISTRIBUTION AT THE 0.10 (*), 0.05 (**), AND 0.01 (***) LEVELS OF SIGNIFICANCE USING THE MODIFIED KOLMOGOROV STATISTIC AND THE PSEUDOCRITICAL VALUES IN STEPHANS (1974) FOR NORMALITY, MEAN AND VARIANCE UNKNOWN.

study suggests that, in the case of sampling from two normal distributions with equal variances, $\hat{D}$ can be regarded as following the folded-normal distribution, with $\xi = D$ for sufficiently large samples, and that, after Elandt (1961), the sampling distribution of $\hat{D}$, and thus that of $\hat{OVL}$, becomes approximately normal when $\mu_f/\sigma_f > 3$.

The failure of the simulated sampling distribution of $\hat{D}$ to follow the folded-normal distribution in the unequal population variance situation is not that surprizing, since the rationalization for the folded-normal $\hat{D}$ is the equal variance formulation of equation 2.16. The failure of the folded-normal model for $\hat{D}$ in the unequal variance case is apparent in the noncentral χ^2 probability plots for the Monte Carlo trials where $\sigma_1^2 \neq \sigma_2^2$. Two such plots are reproduced in figure 2.3 and figure 2.4, illustrating the best and worst fits respectively of the Monte Carlo distribution function of $\hat{D}$ to the folded-normal distribution--as indicated by the Kolmogorov statistic--among the simulation trials where $\sigma_1^2 \neq \sigma_2^2$ and where the folded-normal distribution is rejected. In each probability plot, systematic deviation of the Monte Carlo distribution of $\hat{D}$ from the folded-normal is clear, as the noncentral χ^2 distribution function exceeds the empirical distribution at both low and high ends of the distribution of $\hat{D}^2/\tau^2$ and falls short in between.

An Approximate Confidence Interval for OVL: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Given the problems of bias and estimation of the standard error of OVL evident in the result of the Monte Carlo study, a somewhat different approach may prove more useful in gauging the uncertainties

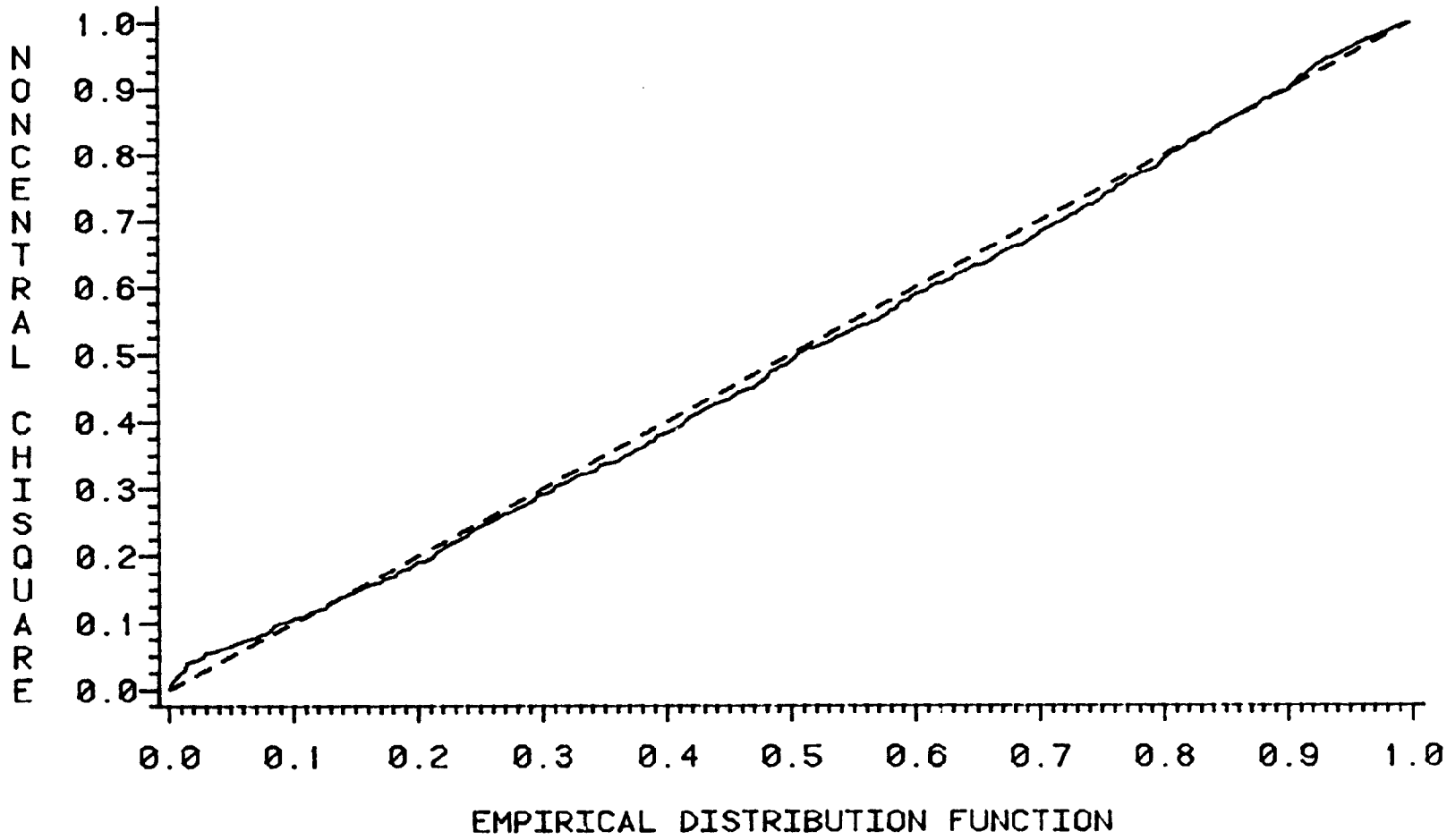

Figure 2.3 Noncentral χ^2 probability plot for $\hat{D}^2/\tau^2$, $\mu_2 = 0.00$, $\sigma_2^2 = 1.2$, and $n_1 = n_2 = 250$. The empirical distribution function is constructed from the 1000 Monte Carlo observations of $\hat{D}$, and the value of the noncentrality parameter is $\lambda = 4.28403$.

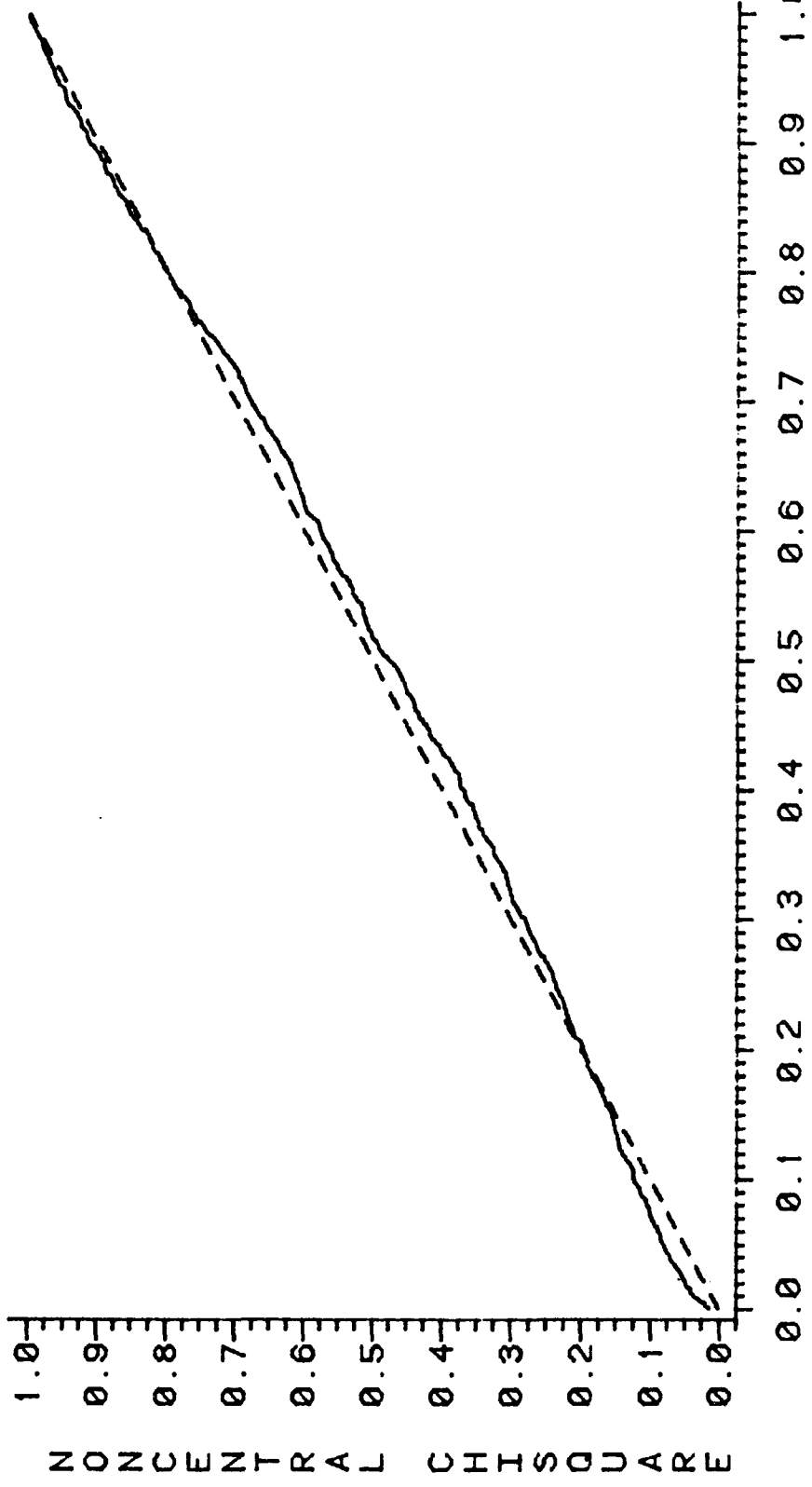

Figure 2.4 Noncentral χ^2 probability plot for $\hat{D}^2/\tau^2$, $\mu_2 = 0.00$, $\sigma_2^2 = 1.2$, and $n_1 = n_2 = 50$. The empirical distribution function is constructed from the 1000 Monte Carlo observations of $\hat{D}$, and the value of the noncentrality parameter is $\lambda = 3.42692$.

inherent to estimation of OVL from sample information. The maximum-likelihood estimator of OVL in the equal variance case can be regarded as a simple transformation of the maximum-likelihood estimator of the Mahalanobis distance separating $f_1(x;\mu_1,\sigma_1^2)$ and $f_2(x;\mu_2,\sigma_2^2)$ to the interval $[0,1]$ through the standard-normal distribution function, $\Phi(z)$. That is, if

$$\hat{\delta}^2 = \left(\frac{\bar{x}_1 - \bar{x}_2}{s_p} \right)^2 , \tag{2.49}$$

then $\hat{OVL} = 2\Phi(-\hat{\delta}/2)$. Moreover (Johnson and Wichern, 1982, p. 468), the random variable F defined by

$$F = \frac{n_1 n_2}{n_1 + n_2} \hat{\delta}^2 \tag{2.50}$$

has a noncentral F distribution with a single numerator degree of freedom, $n_1 + n_2 - 2$ denominator degrees of freedom, and noncentrality parameter λ,

$$\lambda = \frac{n_1 n_2}{n_1 + n_2} \delta^2 = \frac{n_1 n_2}{n_1 + n_2} \left(\frac{\mu_1 - \mu_2}{\sigma} \right)^2 . \tag{2.51}$$

We can use the relationship in equation 2.50 to determine a $(1 - \alpha)100\%$ confidence interval for δ^2 using the appropriate points of the indicated noncentral F distribution, taking as an estimate of the noncentrality parameter

$$\hat{\lambda} = \frac{n_1 n_2}{n_1 + n_2} \hat{\delta}^2 . \tag{2.52}$$

If we denote this interval by (δ_L^2, δ_U^2), a reasonable set of interval boundaries is given by the following:

$$\delta_L^2 = \frac{n_1 + n_2}{n_1 n_2} F\left(\frac{\alpha}{2};\ 1,\ n_1 + n_2 - 2,\ \hat{\lambda}\right) , \tag{2.53}$$

$$\delta_U^2 = \frac{n_1 + n_2}{n_1 n_2} F\left(1 - \frac{\alpha}{2};\ 1,\ n_1 + n_2 - 2,\ \hat{\lambda}\right) . \tag{2.54}$$

The solutions to (2.53) and (2.54) can then be used to obtain a corresponding confidence interval for OVL:

$$\left[2\Phi(-\delta_U/2),\ 2\Phi(-\delta_L/2)\right] \quad . \tag{2.55}$$

The idea of using a confidence interval for δ^2 to construct a confidence interval for OVL follows the argument used by Cheng and Iles (1983) to develop confidence bands for the distribution function of a continuous random variable based on a confidence region obtained for the parameters of its distribution. The confidence interval for OVL is presented here as a proposal, since its properties have not been explored.

(Note that equations 2.49 and 2.50 can be used to derive, by statistical differentials, an alternative approximation formula for the variance of $\hat{OVL}$ to that in equation 2.23. From the Monte Carlo study, however, it appears that this approximation compares unfavorably to that given in equation 2.23 when $|\mu_1 - \mu_2|$ is near or equal to zero, a situation of considerable interest. The noncentral χ^2 distribution is often used for the distribution of F in equation 2.50; see, for example, Anderson, 1958, p. 56. This, of course, is the argument for the use of the noncentral χ^2 distribution as the basis of the sampling distribution of $\hat{OVL}$ in the previous section.)

Discussion

The sampling distribution of the maximum-likelihood estimator of OVL between two normal distributions with equal variances can evidently

be approximated by the folded-normal distribution. The characteristics of the folded-normal distribution account for the behavior of $\hat{OVL}$ observed in the Monte Carlo simulation study: the downward bias of $\hat{OVL}$, the relationship of this bias to OVL, and the approximate normality of the smpling distribution of $\hat{OVL}$ when OVL is sufficiently distant from unity. These characteristics are also exhibited by the maximum-likelihood estimator of OVL in the unequal population variances case, but the sampling distribution of $\hat{OVL}$ in this circumstance cannot be represented by the folded-normal distribution. The utility of $\hat{OVL}$ as an inferential statistic is clouded further by the difficulties associated with the accurate estimation of its variance. For the equal variance case, the approximation to the variance of $\hat{OVL}$ appears to provide accurate estimates of the sampling variance of $\hat{OVL}$ as observed in the Monte Carlo study. When the population variances are unequal, the approximation to the variance of $\hat{OVL}$ based on statistical differentials seems reasonable when OVL is not close to 1.0, and it overstates the sampling variance of $\hat{OVL}$ when OVL is near unity.

Thus the value of OVL as a measure of association between two normal distributions requires that it be based on known distributions of on samples large enough that they can be assumed sufficiently representative of their populations to be considered equivalent to the populations themselves. To this extent, then, use of $\hat{OVL}$ depends on extremely large sample sizes or, alternatively, making the interpretation as well as the computation of $\hat{OVL}$ conditional upon the observed sample outcomes, treating $\hat{OVL}$ as OVL computed from the sample realizations of the parameters of the normal distributions in question. The overlapping

coefficient may still prove useful in such situations, since it demonstrates the level of agreement between two samples which, on the basis of other statistical tests, represent two distinct but closely associated populations. The biased behavior of $\hat{OVL}$, in fact, suggests that the true overlap between two normal populations, particularly when the observed $\hat{OVL}$ is near unity, is probably higher than that indicated by $\hat{OVL}$, a property which may have the desirable effect of tempering an overenthusiastic conclusion based solely on statistical tests of the equality of the parameters of two normal distributions without concern for the magnitude of any differences detected.

An Example

As an example of the use of $\hat{OVL}$, 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). 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 censuses 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 described 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.

As one might expect, the distribution of estimated 1850 wealth is highly skewed. Examination of these data suggests that a logarithmic transformation is most appropriate, and the natural logarithm of estimated wealth in 1850 is reported in table 2.4 for the 317 persistent and the 284 nonpersistent farmers in the reduced sample. Using these natural logarithms, the sample mean for the persistent farmers is 7.570876, and the sample variance, computed according to equation 2.12, is 2.274353. For the nonpersistent farmers, the sample mean is 7.045991, and the sample variance is 2.303979. An F-test for the equality of the population variances yields an F-ratio of 1.0130 ($p = 0.9068$), so equal population variances will be assumed. The usual t-test for equality of population means yields a t-statistic of 4.2396, which, with 599 degrees of freedom, is statistically significant at the 0.0001 level. Thus it appears entirely reasonable to conclude that the mean wealth of persistent Alabama farmers exceeded the mean wealth of their nonpersistent counterparts, indicating that the migration of Alabama farm operators between 1850 and 1860 to some degree selected the poorer farmers.

The degree of selectivity depends not on the difference in population means but instead on the actual difference in the distributions 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. However, if we compute the maximum-likelihood estimate of the common population variance and use equation

TABLE 2.4

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	6.68081	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
6.94724	6.94724	6.97980	6.98770	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
7.53363	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

TABLE 2.4 (CONTINUED)

FARMERS WHO PERSISTED TO 1860 (N = 317)

7.95330	7.95330	7.98781	8.00523	8.00786	8.01839	8.03306	8.04002	8.06254	8.06639
8.08577	8.08577	8.09399	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
9.38209	9.38209	9.39970	9.40017	9.44974	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 1860 (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	5.17668	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

TABLE 2.4 (CONTINUED)

FARMERS WHO DID NOT PERSIST TO 1860 (N = 284)

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
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	6.85595	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
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

TABLE 2.4 (CONTINUED)

FARMERS WHO DID NOT PERSIST TO 1860 (N = 284)

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	9.63225	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						

SOURCE: A SIMPLE RANDOM SAMPLE OF FARM OPERATORS FROM THE 1850 MANUSCRIPT CENSUS OF AGRICULTURE DESCRIBED IN INMAN (1981).

2.16 to calculate $\hat{OVL}$, we obtain $\hat{OVL}$ = 0.859614, which certainly indicates that the distributions of wealth for these two groups of Alabama farmers are not as distinct as a simple comparison of the sample means might suggest. This leads us to conclude that, while the persistent and nonpersistent Alabama farmers differed in mean wealth, the actual difference in the distributions of wealth for these farm operators is not particularly great.

We can use equation 2.25 to compute an estimate of the standard error of $\hat{OVL}$. Here the estimated variance of $\hat{OVL}$ is 0.00104634; thus the estimated standard error of $\hat{OVL}$ is 0.032347. We may also construct a confidence interval for OVL using the estimated Mahalanobis distance, $\hat{\delta}^2$ = 0.120394. From this, we see the estimated noncentrality parameter for the required points of the appropriate noncentral F distribution is $\hat{\lambda}$ = 18.0346. Using the FINV function in SAS (Hardison et al., 1983) with 1 numerator and 599 denominator degrees of freedom, we find that the limits of a 90 percent confidence interval for δ^2 are δ_L^2 = 0.045069 and δ_U^2 = 0.233791. We than obtain the corresponding 90 percent cenfidence interval for OVL using equation 2.55: (0.808967, 0.915465).

All of the computations performed here are based on the assumed normality of the two distributions compared. In this example, normally distributed natural logarithms of wealth imply that the wealth distributions are log-normal. However, Kolmogorov tests for the normality of the natural logarithms of estimated wealth, using the Stephans (1974) modifications and pseudocritical values, indicate that the natural logarithms of estimated 1850 wealth are <u>not</u> normally distributed.

We shall return to this example in the following chapter, where a nonparametric approach for the estimation of OVL is developed.

Chapter Three

NONPARAMETRIC ESTIMATION OF OVL

The calculation and estimation of the overlapping coefficient based on the assumed normal form of the density functions $f_1(x)$ and $f_2(x)$ have been addressed in the previous chapter. The invariance property of OVL noted in Chapter One provides that if some normalizing transformation (Tukey, 1957; Box and Cox, 1964) can be found and applied to both sets of sample data, the machinery and conclusions concerning the estimation of OVL in the normal case can be implemented. As we have seen, the derivation of explicit or implicit formulations of OVL in other distributional settings is also possible. Suppose, however, that either the specific problem of interest or the data gathered to investigate it suggest no reasonable parametric form for $f_1(x)$ and $f_2(x)$ or lead to rejection of the presumed parametric distribution. In such circumstances there are two obvious approaches for the estimation of OVL. One can adopt a "quasi-parametric" approach, using a flexible family of distribution functions, like th Pearson, Burr, or Johnson families of distributions (Johnson and Kotz, 1970, pp. 9-33), to characterize the two distributions and thereby to estimate OVL. The other approach is to estimate the two distributions nonparametrically, using one of several nonparametric density estimation procedures (Wegman, 1972, 1982). The second of these paths is explored here.

The nonparametric method investigated here uses piece-wise polynomial functions to estimate OVL from two independent samples from the unknown distributions $f_1(x)$ and $f_2(x)$. By fitting quadratic spline functions to the empirical distribution functions through weighted least-squares, taking the derivatives of these spline functions as the estimated densities, and using these density estimates to determine points of intersection, it is possible to obtain a nonparametric estimate of OVL between $f_1(x)$ and $f_2(x)$. The bootstrap (Efron, 1982) provides a natural method of obtaining an estimate of the variance of the estimated OVL, $\tilde{OVL}$. Because the estimation of OVL via quadratic splines substitutes a numerical technique for knowledge about the distributions from which the sample data arose, the discussion of the spline-estimator $\tilde{OVL}$ which follows will be more desciptive than mathematical in orientation. To learn something of the properties of $\tilde{OVL}$ as an estimator of OVL, $\tilde{OVL}$ is compared to $\hat{OVL}$ using a subset of the Monte Carlo data generated from two normal distributions introduced in the previous chapter. The Monte Carlo evidence suggests that $\tilde{OVL}$ can perform well as an estimator of OVL. Like $\hat{OVL}$, $\tilde{OVL}$ is a biased estimator of OVL, and, because of this bias, $\tilde{OVL}$ generally underestimates the true overlap between the normal distributions of interest. The bias of $\tilde{OVL}$ is related to OVL and the sizes of the two samples in the same manner as $\hat{OVL}$; but when sampling from normal distributions, the bias of $\tilde{OVL}$ almost always exceeds the bias of $\hat{OVL}$. As we should expect for a nonparametric estimator, the variance of $\tilde{OVL}$ is greater than that of $\hat{OVL}$ when sampling from two normal distributions.

Spline Density Estimation

The use of polynomial splines to estimate the unknown density of a continuous random variable from sample data $x_1,\ldots,x_n$ is one of a number of related nonparametric techniques of density estimation (Wegman, 1972; Wegman, 1982). Introduced and developed in Boneva et al. (1971), de Montricher et al. (1975), and Wahba (1975), the idea behind spline estimates of density functions is quite simple, and the spline density estimator exhibits desirable statistical properties. It is equivalent to the first derivative of a spline fitted to the empirical distribution function. Given a suitable penalty function, the spline-estimated density is the maximum-penalized-likelihood estimator of the unknown density. Statistical properties of spline-estimated densities have been investigated in several situations (see Wegman, 1982, for a brief review and citations). Lii and Rosenblatt (1975) and Rosenblatt (1977) derive the bias, variance-covariance structure, and asymptotic distributional behavior of densities estimated with cubic splines computed with equally-spaced breakpoints, for example.

Let us begin by defining the empirical distribution function $F_n(x)$, computed from the simple random sample $x_1,\ldots,x_n$:

$$F_n(x) = \begin{cases} 0 & \text{if } x < x_{(1)}\ , \\ \frac{i}{n+1} & \text{if } x_{(i)} \le x < x_{(i+1)}\ ,\ i=1,\ldots,n-1, \\ \frac{n}{n+1} & \text{if } x \ge x_{(n)}\ , \end{cases} \tag{3.1}$$

where $x_{(i)}$ represents the i^{th} sample order statistic. (The rationale for the divisor n+1 in equation 3.1, rather than the more usual divisor n, is the expectation of the probability-integral transform of the order statistics $x_{(1)},\ldots,x_{(n)}$; see Gibbons, 1971, pp. 23, 32. For large n, of course, this difference becomes trivial.) Based on the relationship between $F_n(x)$ and the binomial distribution (Gibbons, 1971, pp. 74-75), the variance of $F_n(x)$ as defined in (3.1) is given by

$$\mathrm{Var}\left(F_n(x)\right) = \frac{n \cdot F_n(x)\left[1 - F_n(x)\right]}{(n+1)^2} , \tag{3.2}$$

and thus $F_n(x)$ is a consistent estimator of the unknown distribution function $F(x)$.

To obtain the spline-estimated density, $\hat{f}(x)$, we fit a polynomial spline to $F_n(x)$; designate this piece-wise polynomial function $\hat{F}(x)$, which can be defined, after de Boor (1978), as follows. First, define a strictly increasing sequence of $\ell+1$ points, $t_1 < t_2 < \ldots < t_\ell < t_{\ell+1}$, such that $x_1,\ldots,x_n$ are contained in the interval $[t_1, t_{\ell+1}]$; we ignore for the moment how ℓ and $t_1,\ldots,t_{\ell+1}$ are determined. Now define ℓ polynomials of degree k-1 as follows:

$$P_i(x) = a_{i1} + a_{i2}x + \ldots + a_{ik}x^{k-1} , \quad i=1,\ldots,\ell; \tag{3.3}$$

where the constants a_{ij} ($l=1,\ldots,\ell$; $j=1,\ldots,k$) must be determined. The spline-estimated distribution function is then defined by the following equation.

$$\hat{F}(x) = P_i(x) \ , \quad t_i < x < t_{i+1} \ ; \ i=1,\ldots,\ell. \tag{3.4}$$

The estimated density, $\hat{f}(x)$, is obtained by differentiating $\hat{F}(x)$: that is,

$$\hat{f}(x) = \frac{\partial}{\partial x} P_i(x) = a_{i2} + 2a_{i3}x + \ldots + (k-1)a_{ik}x^{k-2} \ ,$$
$$t_i < x < t_{i+1} \ ; \ i=1,\ldots,\ell. \tag{3.5}$$

The properties of $F(x)$ and $f(x)$ require natural constraints on $\hat{F}(x)$ and $\hat{f}(x)$, and these restrictions are incorporated into the computation of $\hat{F}(x)$ as constraints on the constants a_{ij} ($i=1,\ldots,\ell$; $j=1,\ldots,k$). The continuity of $\hat{F}(x)$ can be assured by imposing the conditions

$$P_i(t_{i+1}) = P_{i+1}(t_{i+1}) \ , \ i=1,\ldots,\ell-1. \tag{3.6}$$

Similar conditions make $\hat{f}(x)$ continuous as well:

$$\frac{\partial}{\partial x} P_i(t_{i+1}) = \frac{\partial}{\partial x} P_{i+1}(t_{i+1}) \ , \ i=1,\ldots,\ell-1. \tag{3.7}$$

In addition, we should insist that

$$\hat{F}(t_1) = 0 \ , \tag{3.8}$$

$$\hat{F}(t_{\ell+1}) = 1 \ , \tag{3.9}$$

and

$$\hat{f}(x) = \frac{\partial}{\partial x} P_i(x) \geq 0 \ , \ t_1 \leq x \leq t_{\ell+1} \ ; \ i=1,\ldots,\ell. \tag{3.10}$$

Finally, it will prove convenient to let

$$\hat{f}(t_1) = \hat{f}(t_{\ell+1}) = 0 \ . \tag{3.11}$$

$\hat{F}(x)$, and thus $\hat{f}(x)$, is obtained from $F_n(x)$ by weighted least-squares, subject to the constraints in equations 3.6 through 3.11, to determine the spline coefficients a_{ij} (i=1,...,ℓ; j=1,...,k). (For the least-squares approach to the general use of splines, see Wold, 1974; Buse and Lim, 1977; Suits et al., 1978; Sampson, 1979; Smith, 1979; and Wegman and Wright, 1983.)

Here the FORTRAN routine FC written by Hanson (1979) is used to specify the constraints in equations 3.6 through 3.11 and, following de Boor (1978), to compute the coefficients of the quadratic spline fitted to $F_n(x)$ by weighted least-squares, using the variance estimate of $F_n(x)$ at each data point $x_1,...,x_n$ from equation 3.2 to determine the appropriate weights. The spline coefficients obtained from FC follow de Boor's basis-spline, or B-spline, notation rather than the piece-wise polynomial format of (3.4), and the FORTRAN routine BVALUE (de Boor, 1978) can be used to evaluate the spline-estimate of the distribution function, $\hat{F}(x)$, and its derivative, $\hat{f}(x)$, at any point desired. Of course, the order of the spline (k), and hence the degree of the polynomials (k-1), and the the sequence of breakpoints $t_1,..., t_{\ell+1}$ must still be specified. But, given two samples, one can construct estimates of the distribution functions, $\hat{F}_1(x)$ and $\hat{F}_2(x)$, and from them estimates of the densities, $\hat{f}_1(x)$ and $\hat{f}_2(x)$, and use them to estimate OVL.

In the investigation of the usefulness of spline-estimated densities in the estimation of OVL presented here, the order of the splines used to construct $\hat{F}_1(x)$ and $\hat{F}_2(x)$ is limited to the case of k = 3; that is, the splines consist of piece-wise quadratic polynomials.

This means that $\hat{f}_1(x)$ and $\hat{f}_2(x)$ will be piece-wise linear functions with the appearance of frequency polygons terminating at the endpoints of the interval $[t_1, t_{\ell+1}]$ with vertices at all interior breakpoints. This stategy attains two objectives. First, nonnegative density estimates on the entire interval $[t_1, t_{\ell+1}]$ are obtained by specifying constraint 3.10 at all interior breakpoints:

$$\hat{f}(t_i) = a_{i2} + a_{i3} t_i \geq 0 \ , \ i=2,\ldots,\ell. \tag{3.12}$$

The problem of negative density estimates using cubic (or higher order) polynomial splines is not academic, and Boneva et al. (1971, pp. 3-4) expressly permit such negative densities in their approach. Second, the calculation of the points of intersection of the estimated densities becomes quite straight-forward when they are piece-wise linear functions.

The specification of the breakpoint sequence $t_1,\ldots,t_{\ell+1}$ really involves three separate issues: the number of subintervals ℓ into which the interval $[t_1, t_{\ell+1}]$ is divided by the breakpoint sequence; the endpoints t_1 and $t_{\ell+1}$; and the determination of the remaining breakpoints, $t_2,\ldots,t_\ell$, in the interval $[t_1, t_{\ell+1}]$.

The number of intervals required for the spline-estimation of an unknown density f(x) is a question with no clearcut answer, and the solution adopted here may strike the reader as somewhat arbitrary. The problem is that while specifying too few subintervals introduces error

stemming from the failure of the quadratic pieces of the spline to fit $F_n(x)$ adequately, specifying too many subintervals creates difficulties of another sort. Because the coefficients of the quadratic terms in the polynomial pieces determine the slopes of the line segments which compose the estimated density, the estimated density will become increasingly erratic as more subintervals are specified and as $F_n(x)$ can be approximated more reasonably by linear terms alone on the more numerous, smaller subintervals. The number of subintervals used here is calculated from the rule proposed in Sturges (1926) for the number of classes in a frequency histogram, rounding down to the nearest interger to obtain ℓ:

$$\ell = 1 + 3.322 \cdot \log_{10}(n') \ , \tag{3.13}$$

where n' is the number of unique points in the sample distribution function $F_n(x)$. The usual criticism of Sturges's rule, that it produces too few histogram classes when the underlying distribution is asymmetric or the sample contains outlying values (Snee and Pfeifer, 1983), does not appear compelling in the role assigned to it here, since, as described below, the values of the interior breakpoints defining the boundaries of the subintervals in $[t_1, t_{\ell+1}]$ are chosen to allow efficient use of the breakpoints rather than to divide $[t_1, t_{\ell+1}]$ into subintervals of equal length. Further investigation may suggest a better algorithm for computing the number of polynomial pieces in the spline

fitted to $F_n(x)$ when, as here, these subintervals are of unequal length. If the breakpoints are chosen so that $t_1,\ldots,t_{\ell+1}$ are equally-spaced, the procedure proposed in Wahba (1975) may be employed to determine the number of subintervals and their common length.

The method used to fix t_1 and $t_{\ell+1}$ is a simple one, based on the transformation of the sample observations to the interval $[0,1]$. Three such transformations are the following. If the domain of the distribution presumed to generate the sample data is the interval $[a,b]$, a simple linear transformation,

$$g(x) = \frac{x - a}{b - a}, \tag{3.14}$$

appears obvious. When the domain of the distribution is assumed to be $[0,\infty)$, we may choose

$$g(x) = \frac{x}{1 + x}, \tag{3.15}$$

Finally, if the domain of the distribution giving rise to the sample is $(-\infty,+\infty)$, then we can use

$$g(x) = \frac{\exp(x)}{1 + \exp(x)} . \qquad (3.16)$$

The spline is then fit to the empirical distribution function on the transformed scale, $F_n(g(x))$, and t_1 and $t_{\ell+1}$ can be set to zero and one respectively. The inverse transformation, $x = g^{-1}(y)$ can then be used to obtain $\hat{F}(x)$, although this is not necessary for the calculation of $\tilde{OVL}$ if the same transformation is applied to both sets of sample data, for OVL, as we have seen, is invariant under such transformation.

Thus only the placement of the $\ell-1$ interior breakpoints, $0 < t_2 < \ldots < t_\ell < 1$, remains. Here the sequence of breakpoints is determined iteratively by fitting the quadratic spline to $F_n(g(x))$ using an initial sequence of breakpoints derived from the empirical distribution function itself; generating a new sequence of breakpoints from the fitted spline with de Boor's NEWNOT routine (de Boor, 1978, pp. 184-86); then recomputing the spline approximation $\hat{F}(g(x))$ with this new breakpoint sequence. The placement of the interior breakpoints generated by NEWNOT appeals to the desideratum that the intervals between the knots, or breakpoints, should be relatively small where $\hat{F}(g(x))$, hence $F_n(g(x))$, is changing rapidly and relatively large where $\hat{F}(g(x))$ is changing slowly. Since the estimated density (on the transformed scale) is the derivative of $\hat{F}(g(x))$, NEWNOT in effect picks the vertices of a frequency polygon representing $f(g(x))$ by shortening and lengthening the intervals between breakpoints, based on the behavior of $F_n(g(x))$. Successive computation of $\hat{F}(g(x))$ and the construction

of new sequences of breakpoints can be continued indefinitely, but the major improvement, indicated by a reduction in the residual sum of squares for the fit of $\hat{F}(g(x))$ to $F_n(g(x))$, appears in the first iteration. The initial sequence of breakpoints used here utilizes equally-spaced quantiles from the empirical distribution function for the interior breakpoints. That is, if equation 3.13 requires ten subintervals ($\ell = 10$), then the deciles from $F_n(g(x))$ are used for the values of the breakpoints between $t_1 = 0$ and $t_{\ell+1} = 1$.

Because the Hanson routine FC used to fit the spline to $F_n(g(x))$ uses the de Boor B-spline representation, additional points, or knots, must be specified outside the interval $[t_1, t_{\ell+1}]$: k-1 points equal to or less than t_1 and k-1 points equal to or greater than $t_{\ell+1}$ (Hanson, 1979, pp. 8-10). The sequence of the k-1 points less than or equal to t_1, the $\ell+1$ points in the interval $[t_1, t_{\ell+1}]$, and the k-1 points greater than or equal to $t_{\ell+1}$ define the knot-sequence of the spline. Following Kozak (1980), the left-hand exterior knots are always set equal to $t_1 = 0$, and the right-hand exterior knots are always set equal to $t_{\ell+1} = 1$. Since here k = 3, the knot-sequence used to obtain $\hat{F}(g(x))$ is $0,0,0,t_2,\ldots,t_\ell,1,1,1$, with $t_2,\ldots,t_\ell$ changing with each pass through NEWNOT.

The procedure adopted here to derive spline estimates of an unknown density can be summarized as follows. First, an appropriate transformation is used to map the sample data to the interval $[0,1]$. Next, the empirical distribution function is constructed from the sample on the transformed scale. The number of quadratic pieces in the spline is calculated from (3.13), and the initial sequence of internal

breakpoints fixed from $F_n(g(x))$. A quadratic spline is fit to $F_n(g(x))$, suject to the conditions embodied in equations 3.6, 3.7, 3.8, 3.9, 3.11, and 3.12, using weighted least-squares. NEWNOT is used to generate a new breakpoint sequence, and the spline is recomputed. The quadratic spline obtained after two passes through the NEWNOT-spline-computation process is taken as $\hat{F}(g(x))$ and its derivative as $\hat{f}(g(x))$. The entire procedure is illustrated in figures 3.1 through 3.10. The sample distribution function computed from a sample of 100 standard-normal deviates generated from the IMSL routine GGNML (IMSL, 1982) and transformed by equation 3.16 is shown in figure 3.1. From equation 3.13, seven quadratic pieces are used in the spline fit to the empirical distribution function, and figure 3.2 illustrates the algorithm described above for obtaining the six breakpoints in the interval $[0,1]$. The fitted spline, using these breakpoints, is shown in figure 3.3, and the derivative of this spline is depicted in figure 3.4. Figure 3.5 shows the fitted spline obtained with the set of breakpoints constructed from the first spline with the NEWNOT procedure; movement of the internal breakpoints is clearly visible when this figure is compared to figure 3.3. The derivative of this quadratic spline is shown in figure 3.6. Figures 3.7 and 3.8 illustrate the spline-estimated distribution function $\hat{F}(g(x))$ and the estimated density $\hat{f}(g(x))$ obtained after a second pass through NEWNOT. Using the inverse of transformation 3.16, one can obtain $\hat{F}(x)$, shown in figure 3.9 with the actual standard-normal distribution function for reference. The estimated density $\hat{f}(x)$ can also be obtained from the inverse transformation, scaling by the appropriate differential, and it is depicted in figure 3.10, together with the standard-normal probability density function.

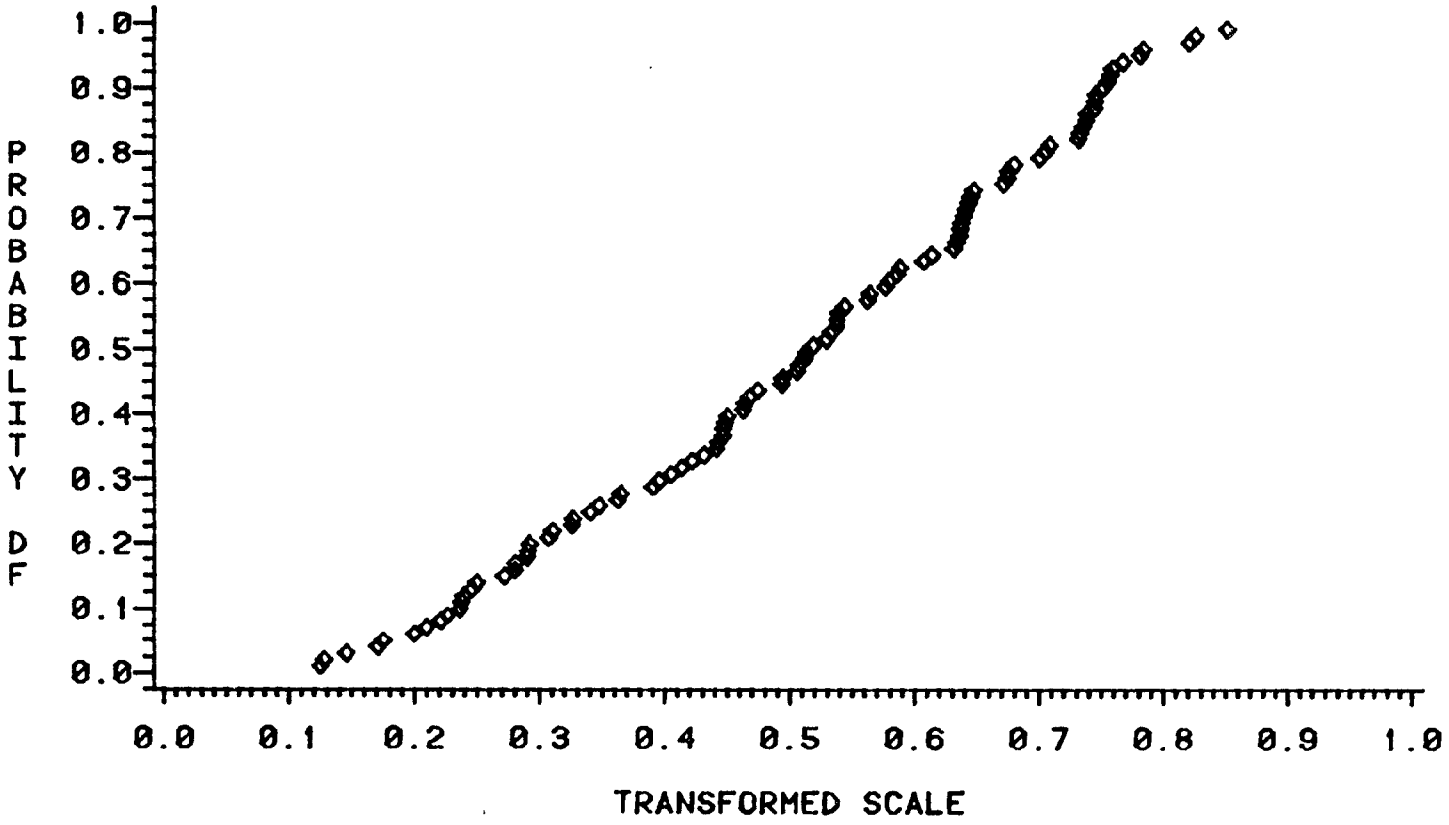

Figure 3.1 Spline density estimation: construction of the empirical distribution function on the transformed scale. This sample distribution function is constructed from a generated sample of 100 standard-normal deviates transformed by equation 3.16.

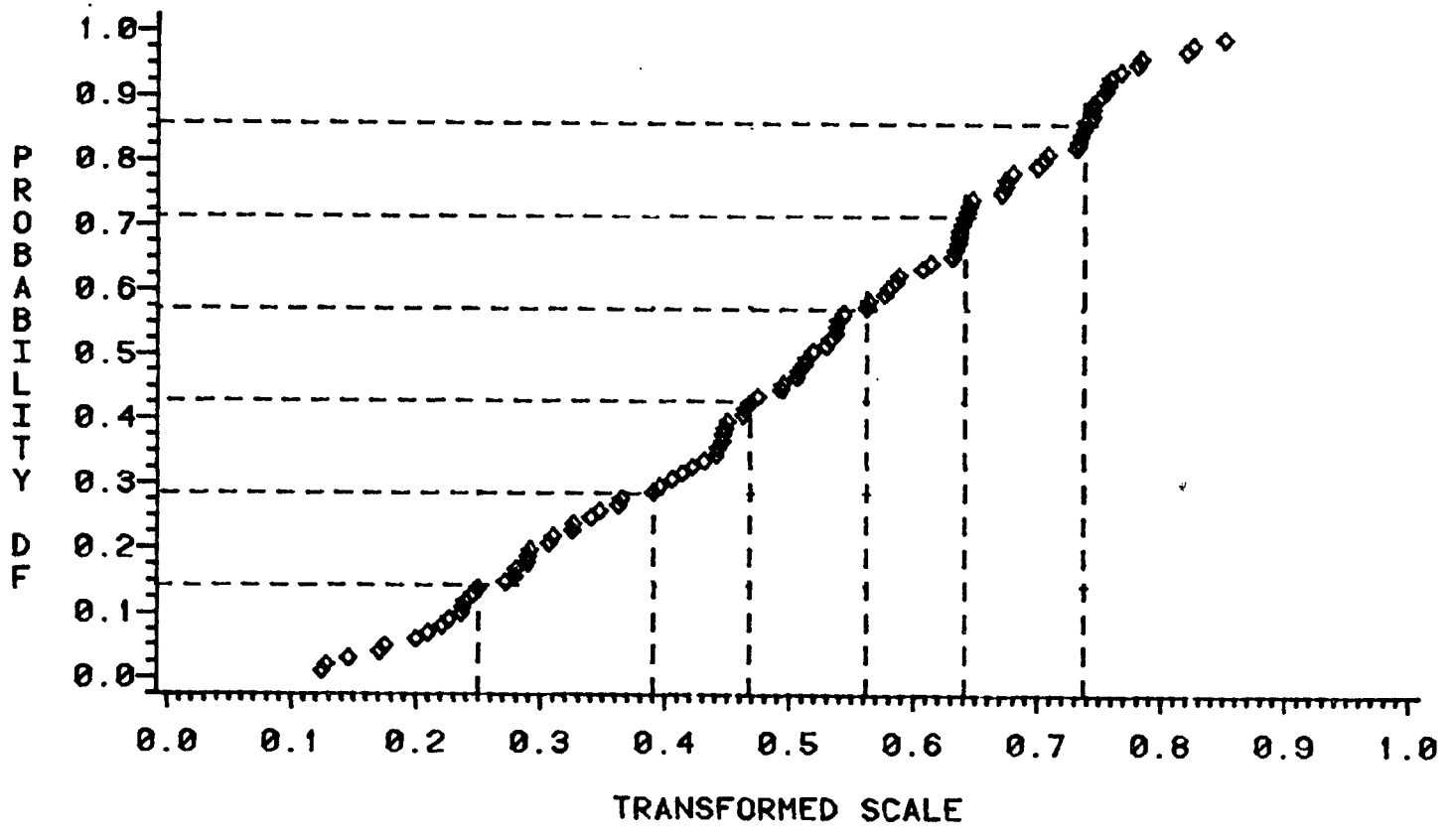

Figure 3.2 Spline density estimation: the algorithm used to construct the initial sequence of interior breakpoints. The horizontal broken lines indicate the equal spacing of these points on the vertical (cumulative probability) scale, while the vertical broken lines indicate the actual breakpoint values on the horizontal axis.

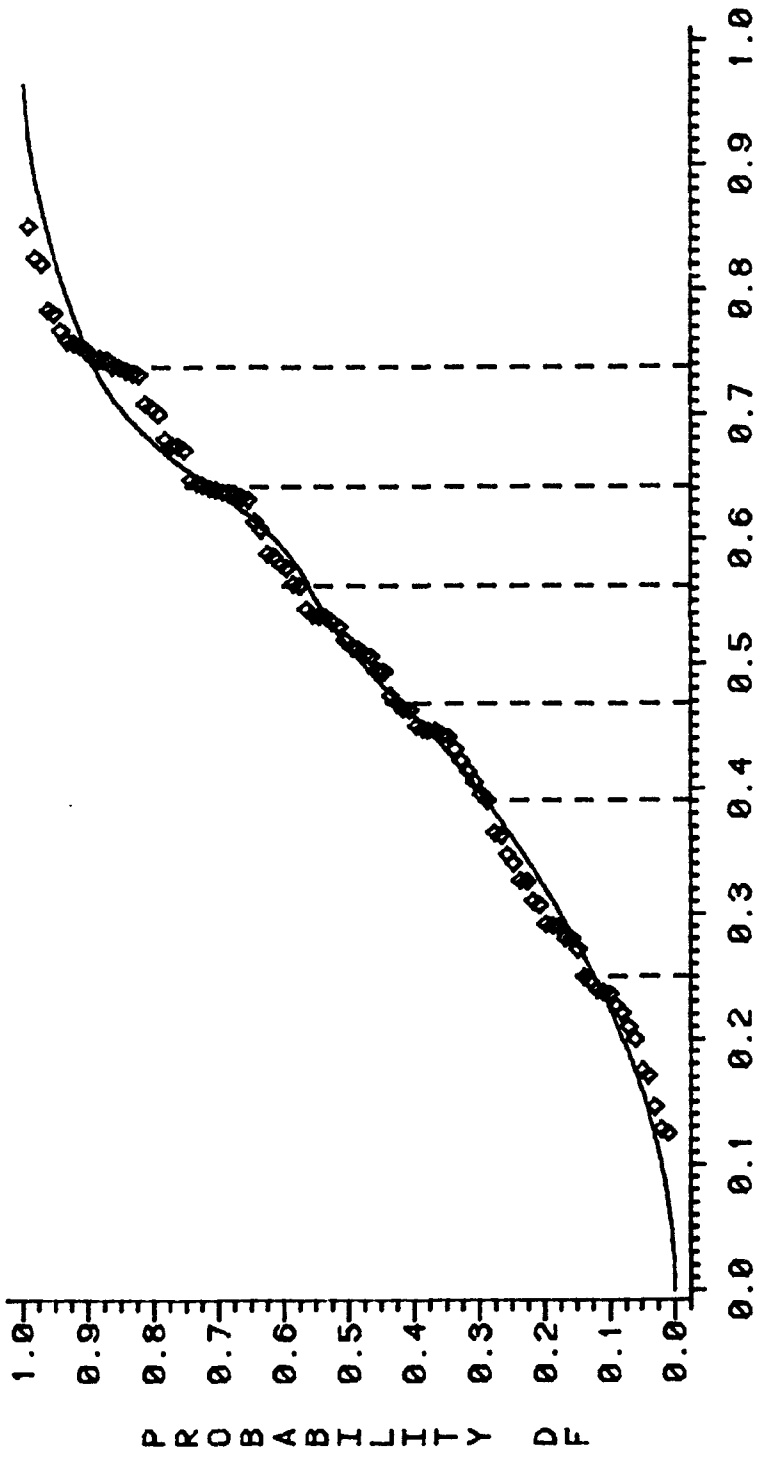

Figure 3.3 Spline density estimation: the spline fitted to the empirical distribution function using the initial breakpoint sequence. The solid line indicates the fitted spline, and the broken vertical lines indicate the initial interior breakpoints.

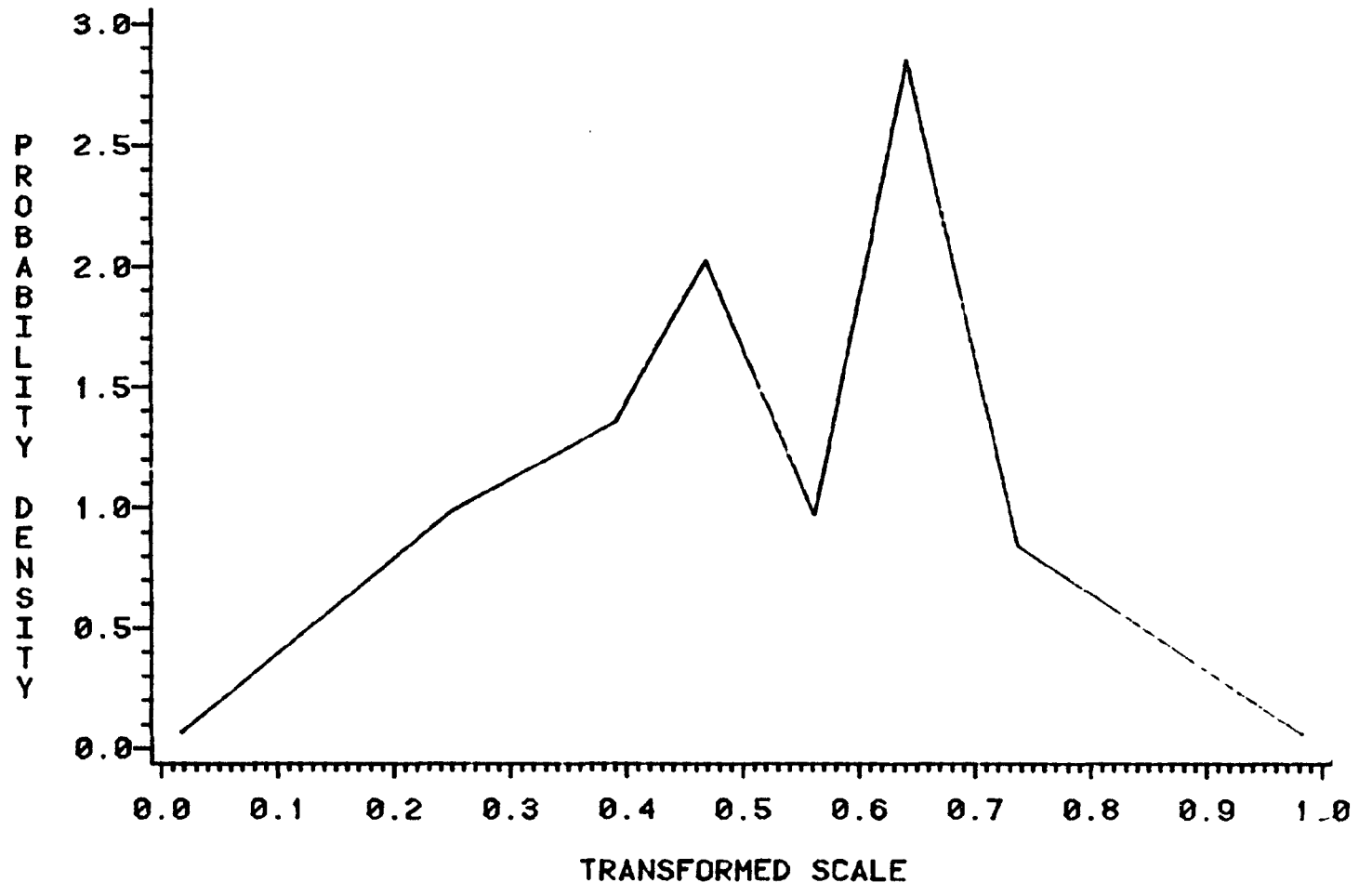

Figure 3.4 Spline density estimation: the estimated density obtained from the spline fitted to the empirical distribution function using the initial breakpoint sequence.

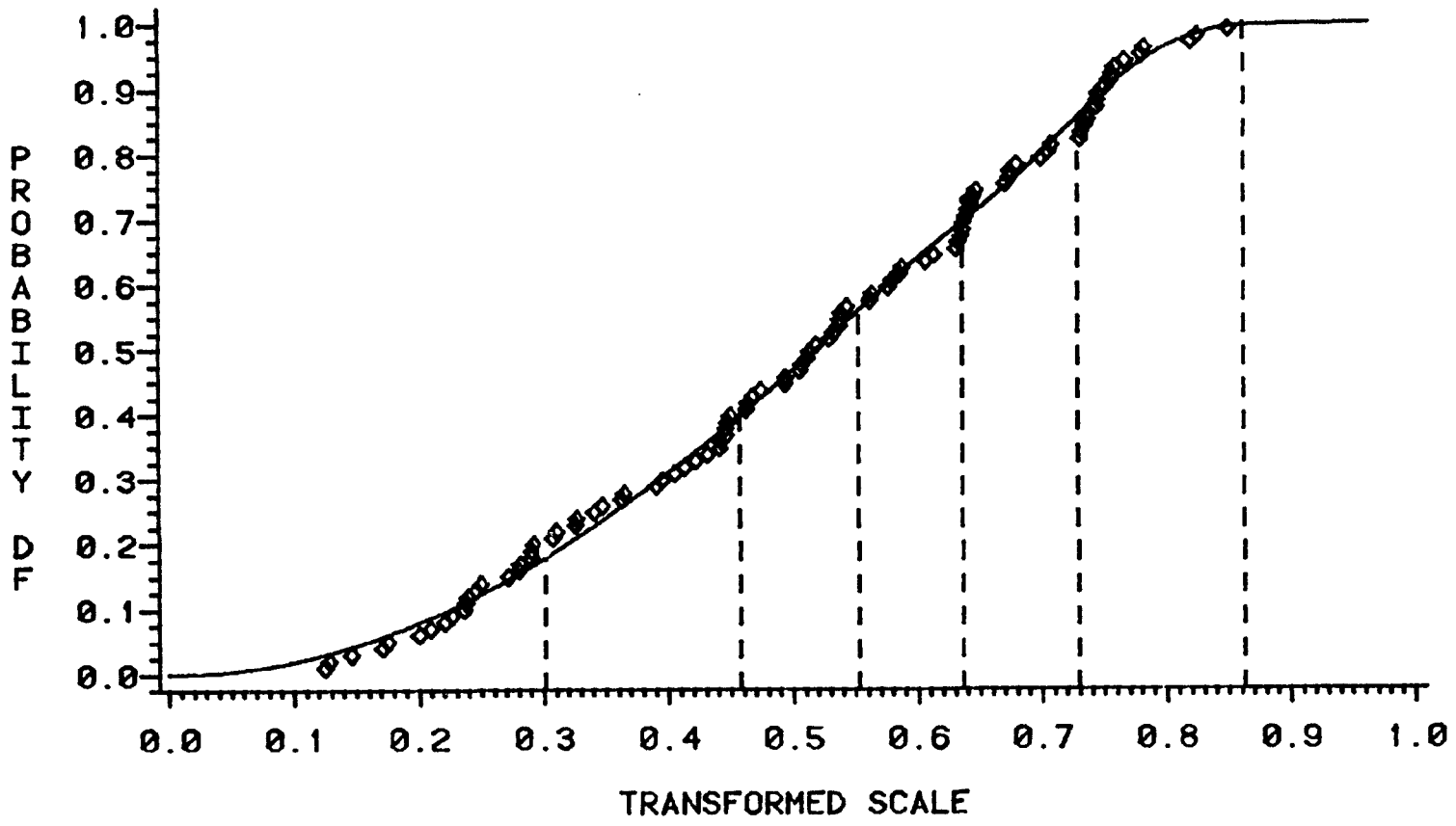

Figure 3.5 Spline density estimation: the spline fitted to the empirical distribution function after the first pass through NEWNOT. The solid line indicates the fitted spline, and the vertical broken lines depict the interior breakpoints obtained from NEWNOT and the spline shown in figure 3.3.

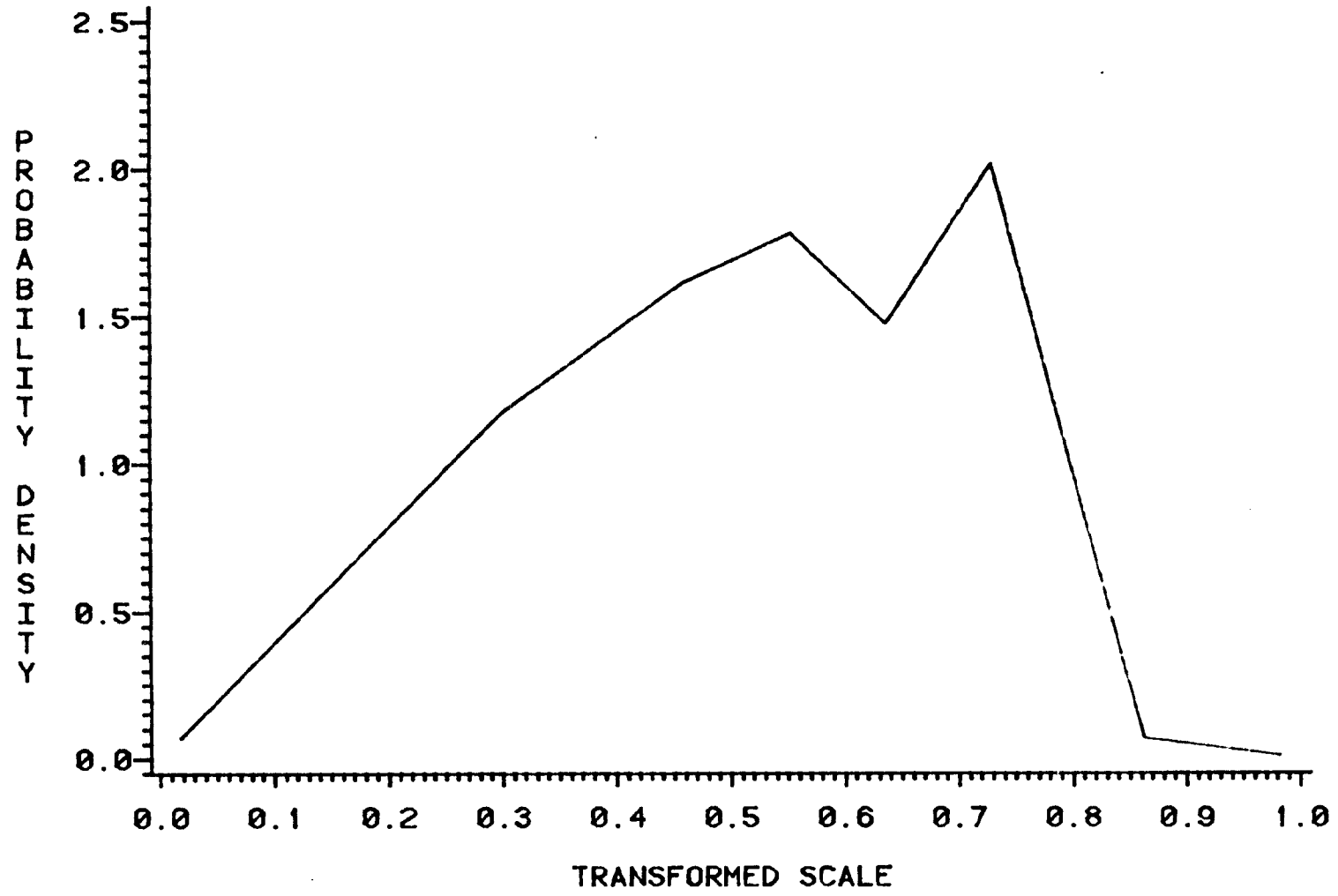

Figure 3.6 Spline density estimation: the estimated density obtained from the spline fitted to the empirical distribution function after the first pass through NEWNOT.

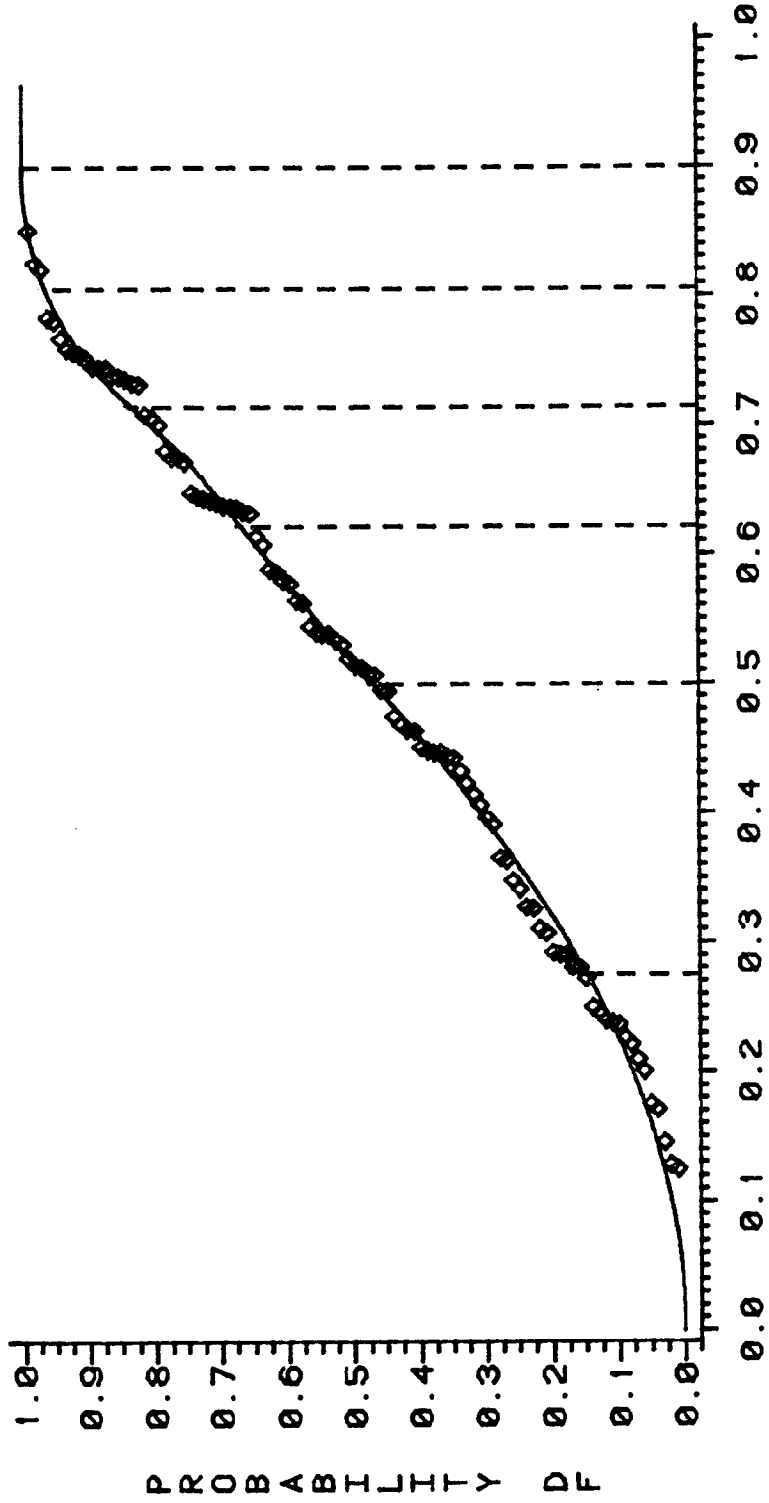

Figure 3.7 Spline density estimation: the spline fitted to the empirical distribution function after the second pass through NEWNOT. The solid line indicates the fitted spline, and the vertical broken lines depict the interior breakpoints obtained from NEWNOT and the spline shown in figure 3.5.

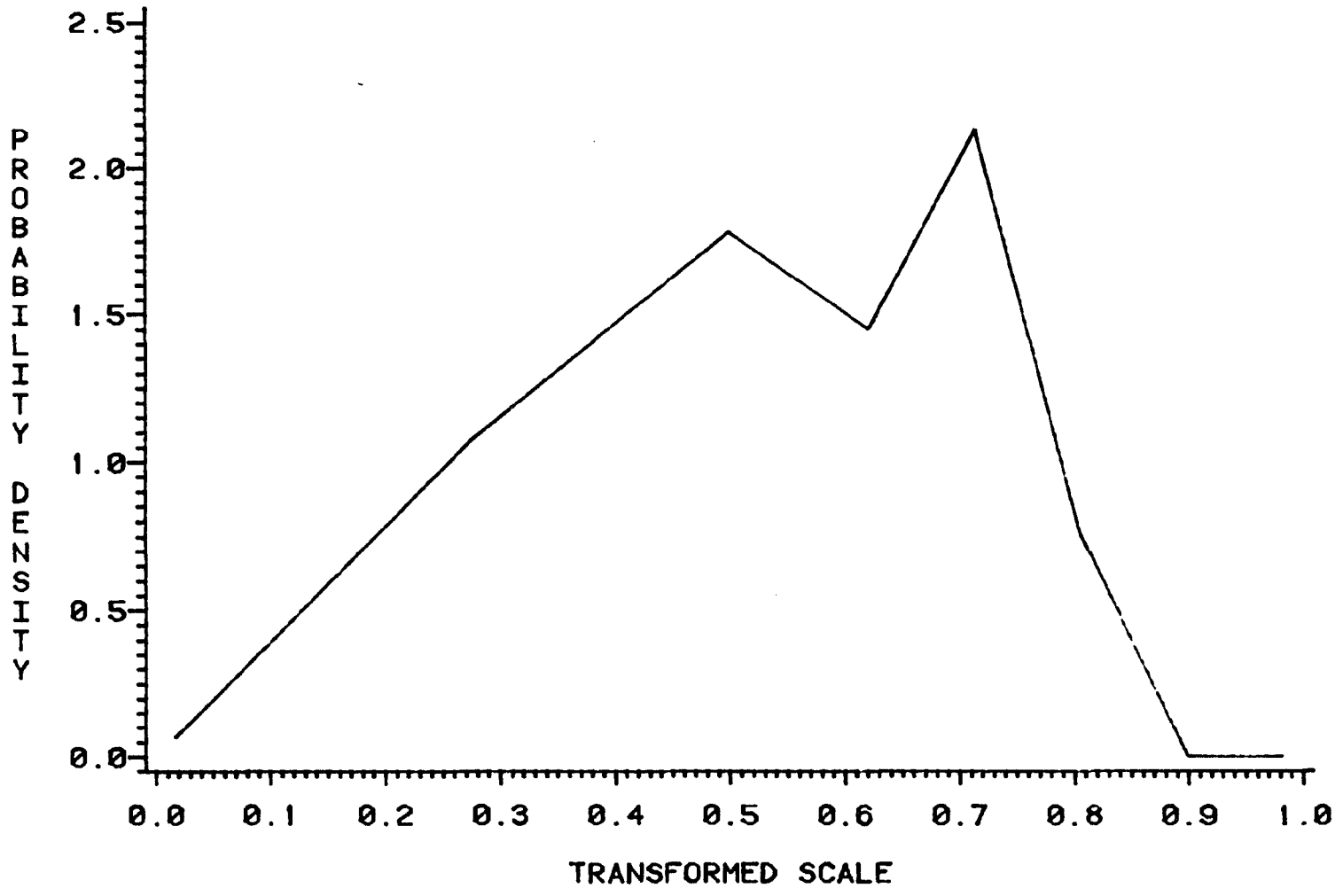

Figure 3.8 Spline density estimation: the estimated density obtained from the spline fitted to the empirical distribution function after the second pass through NEWNOT.

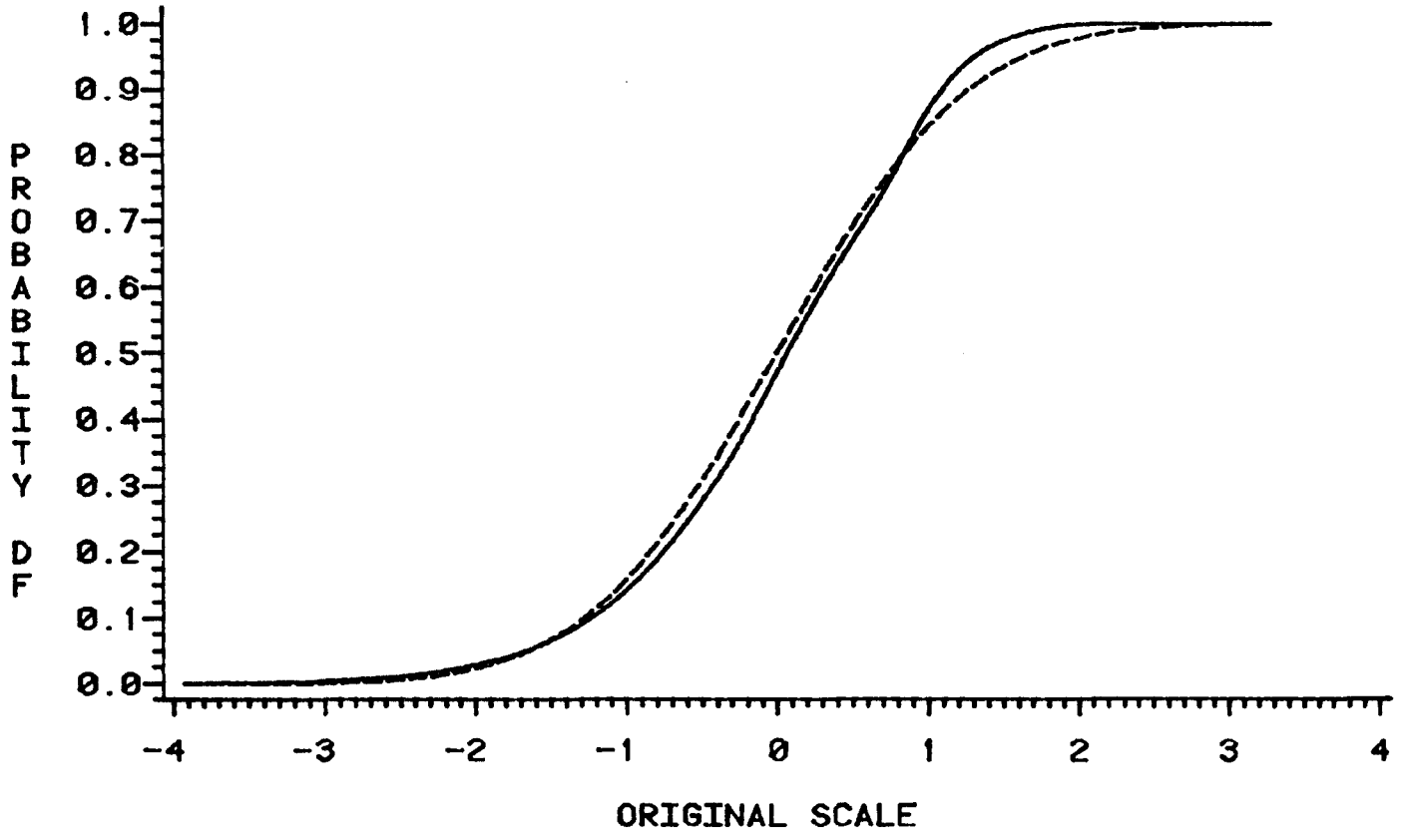

Figure 3.9 Spline density estimation: the spline-estimated distribution function transformed to the original scale. The solid line depicts the estimated distribution function, and the broken line shows the standard-normal distribution function, $\Phi(z)$.

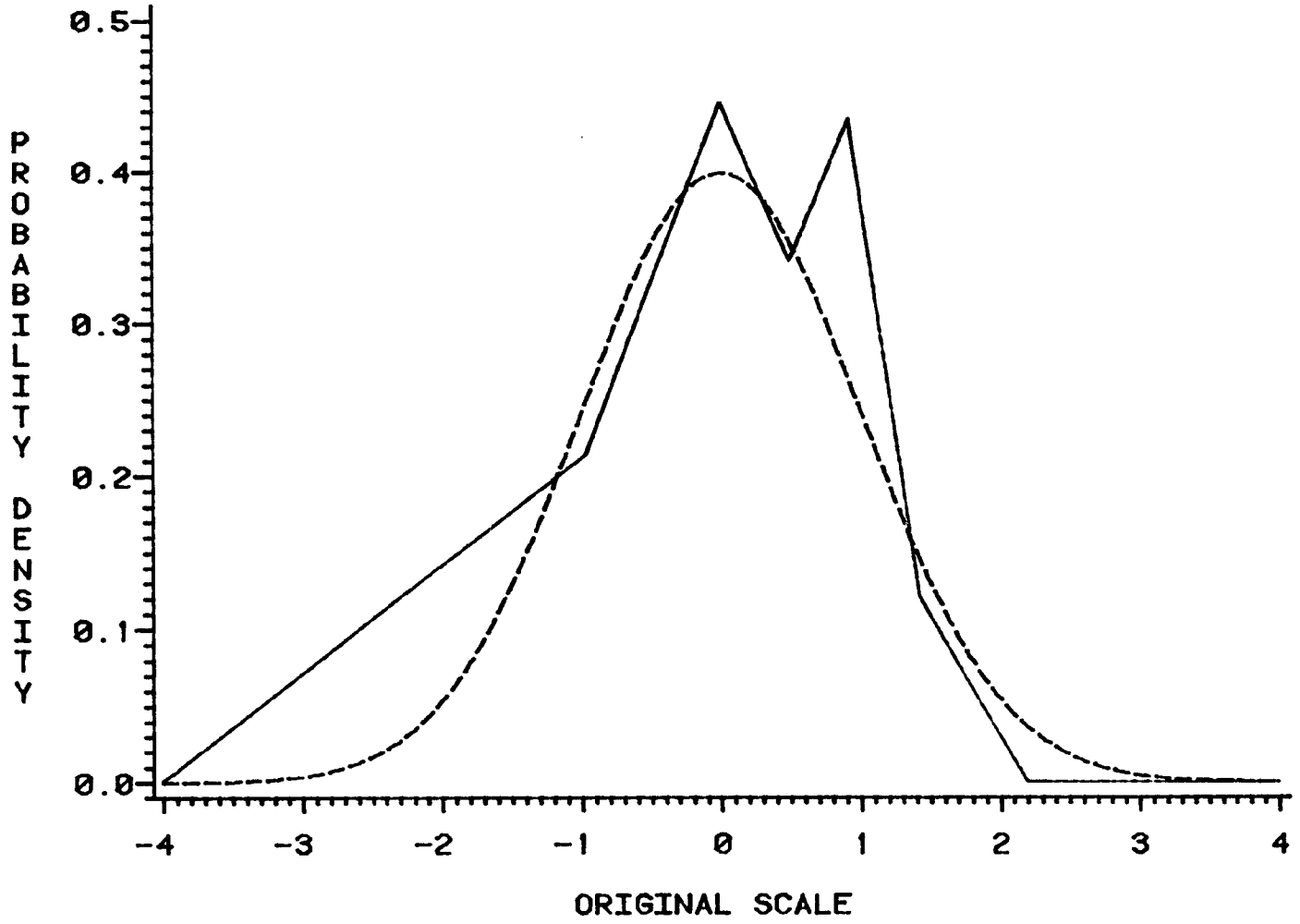

Figure 3.10 Spline density estimation: the spline-estimated density transformed to the original scale. The estimated density is shown by the solid line, and the standard-normal density function by the broken line.

The effect of the number of quadratic pieces chosen to determine $\hat{F}(g(x))$ on the estimated density $\hat{f}(g(x))$ may be seen in figures 3.11 and 3.12. In figure 3.11, five quadratic pieces were used, with the result that the "notch" at the peak of the density estimate based on seven quadratic pieces (figure 3.8) is no longer present. Using nine quadratic pieces introduces additional jaggedness into the density estimate, as figure 3.12 demonstrates. Thus one might conclude that the choice of one or two fewer subintervals than the value of ℓ given by the algorithm based on Sturges's rule seems reasonable here, but the use of any more subintervals is probably unwise. The usefulness of the algorithm and the improvement in the estimated densities which accompany an increase in the sample size become evident in figures 3.13 and 3.14, which show $\hat{F}(x)$ and $\hat{f}(x)$ obtained from a sample of 500 standard-normal deviates, and figures 3.15 and 3.16, which display the spline-estimated distribution function and density based on a sample of size 1000 generated from the standard-normal distribution. These figures demonstrate that the spline-based approach to density estimation outlined here requires very substantial sample sizes for the successful representation--measured only in qualitative terms--of an unknown density.

Estimation of OVL with Quadratic Splines

Given the procedure for estimating an unknown density developed above, obtaining an estimate of OVL based on quadratic splines is remarkably straight-forward. From two independent samples from the two unknown distributions, $x_{11},\ldots,x_{1n_1}$ and $x_{21},\ldots,x_{2n_2}$, we construct the two sample distribution functions, $F_{n_1}(g_1(x))$ and $F_{n_2}(g_2(x))$, and

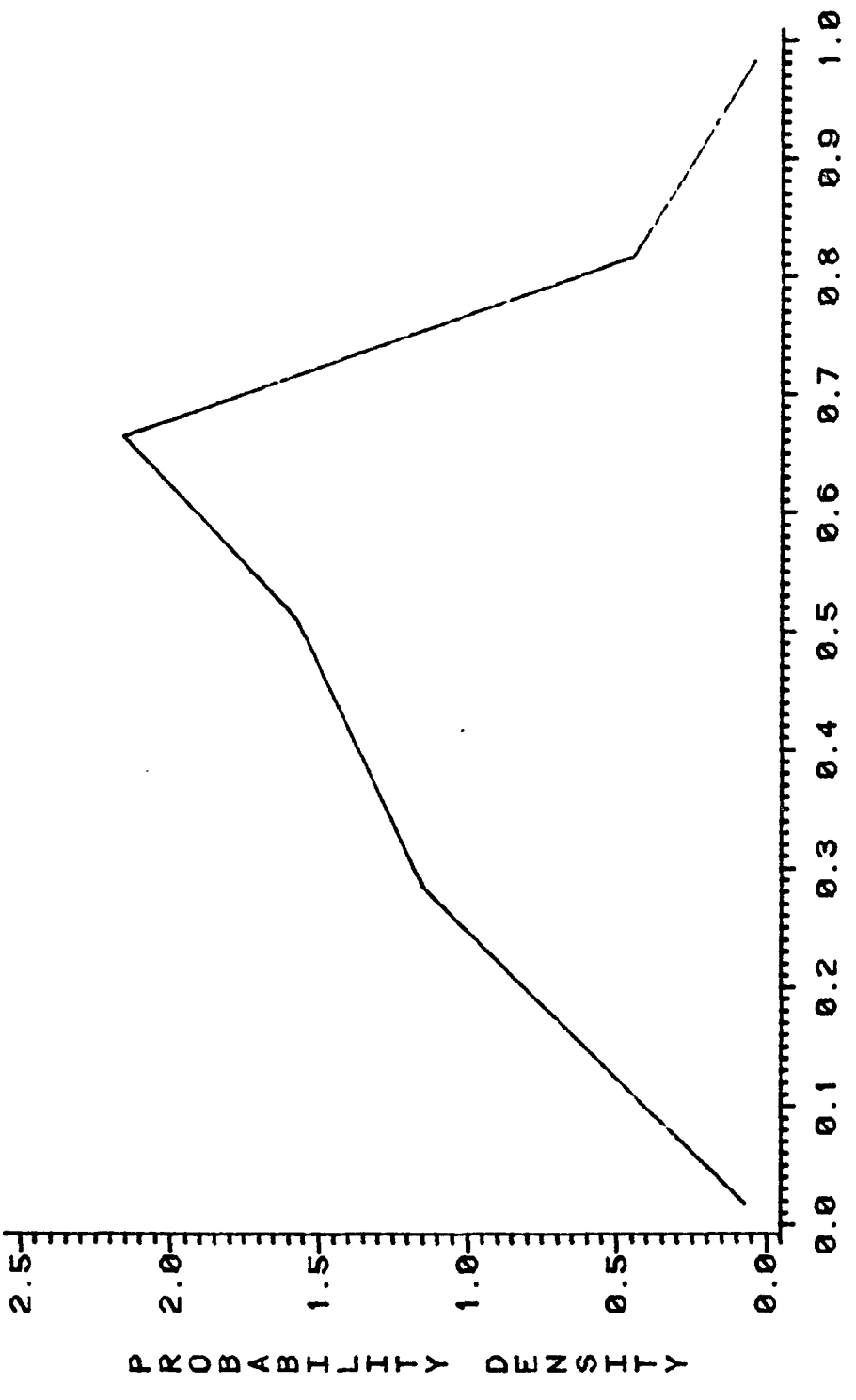

Figure 3.11 Spline density estimation: the estimated density on the transformed scale obtained after two passes through NEWNOT with $\lambda = 5$.

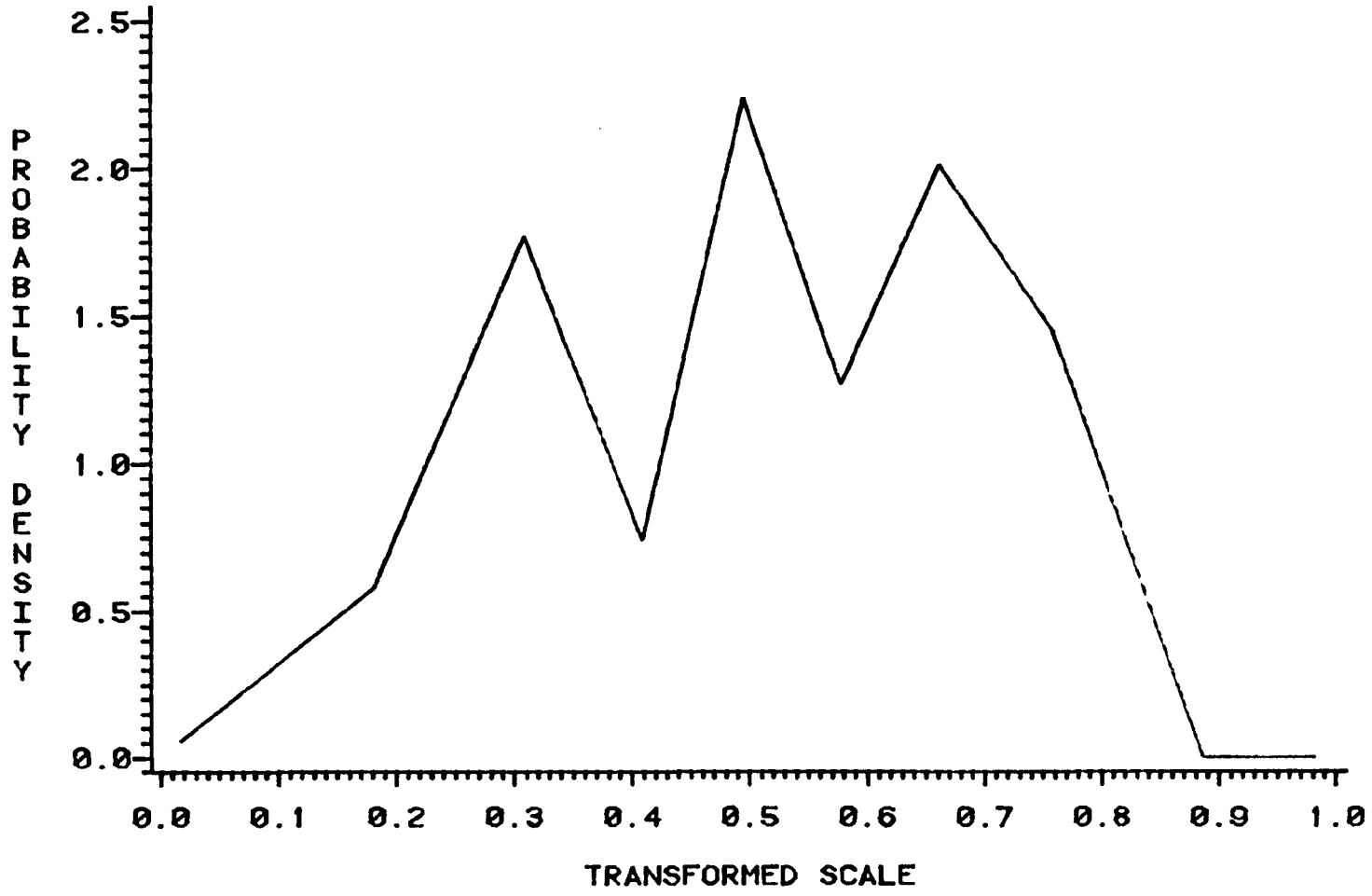

Figure 3.12 Spline density estimation: the estimated density on the transformed scale obtained after two passes through NEWNOT with $\ell = 9$.

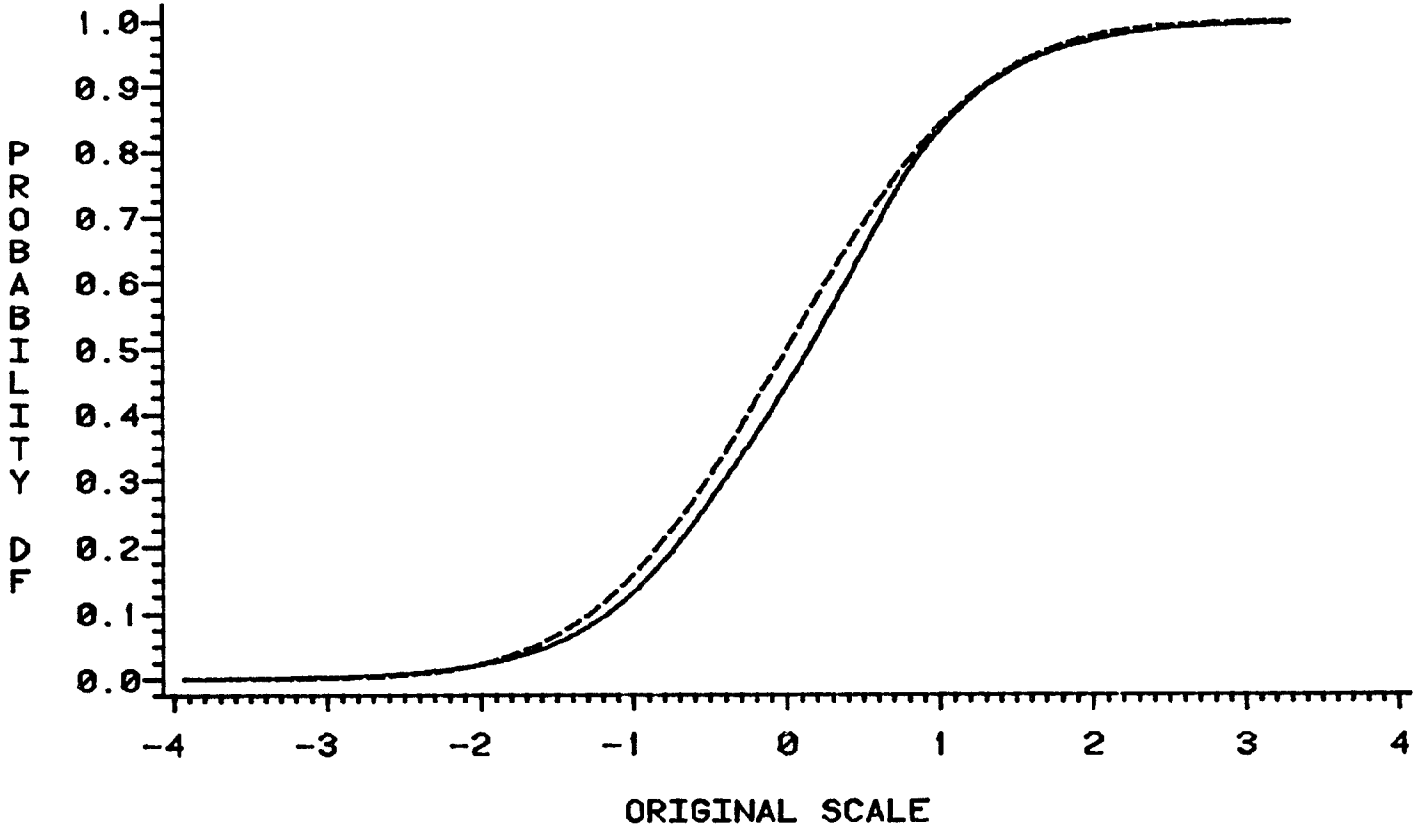

Figure 3.13 Spline density estimation: the spline-estimated distribution function obtained from a generated sample of 500 standard-normal deviates. The estimated distribution function is shown by the solid line, and the standard-normal distribution function $\Phi(z)$ is indicated by the broken line.

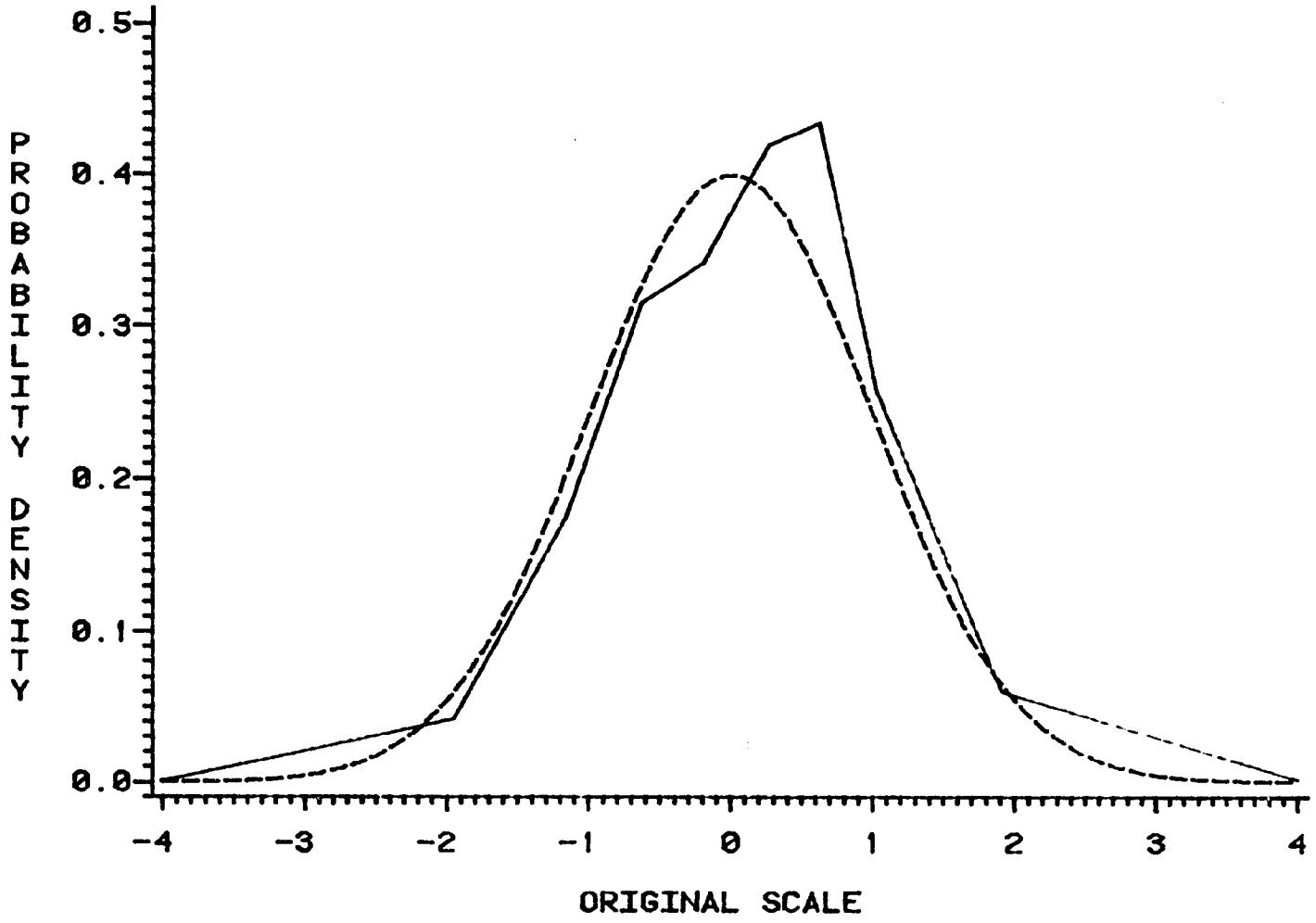

Figure 3.14 Spline density estimation: the spline-estimated density function obtained from a generated sample of 500 standard-normal deviates. The estimated density is shown by the solid line, and the standard-normal density function is indicated by the broken line.

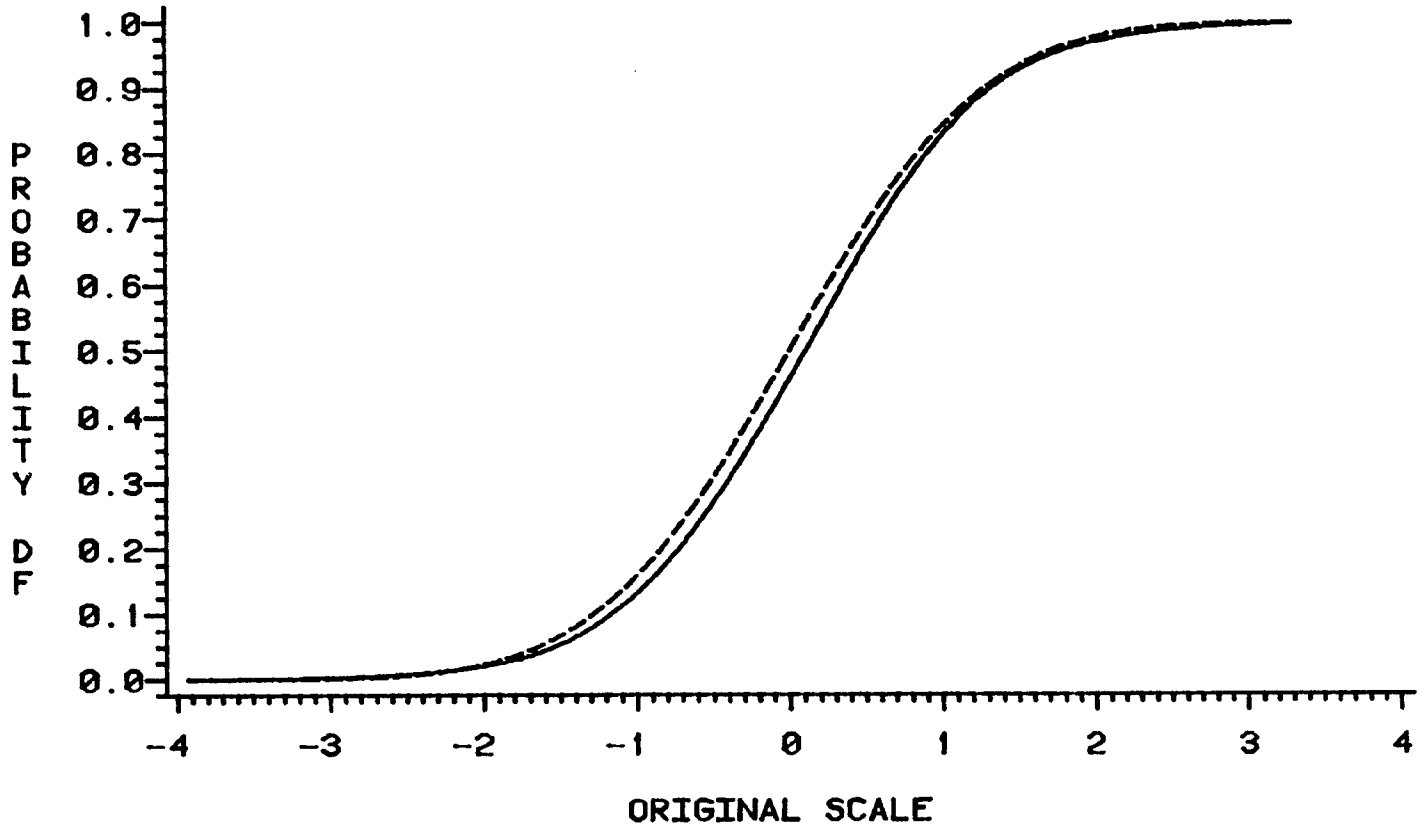

Figure 3.15 Spline density estimation: the spline-estimated distribution function obtained from a generated sample of 1000 standard-normal deviates. The estimated distribution function is indicated by the solid line, and the standard-normal distribution function $\Phi(z)$ is shown by the broken line.

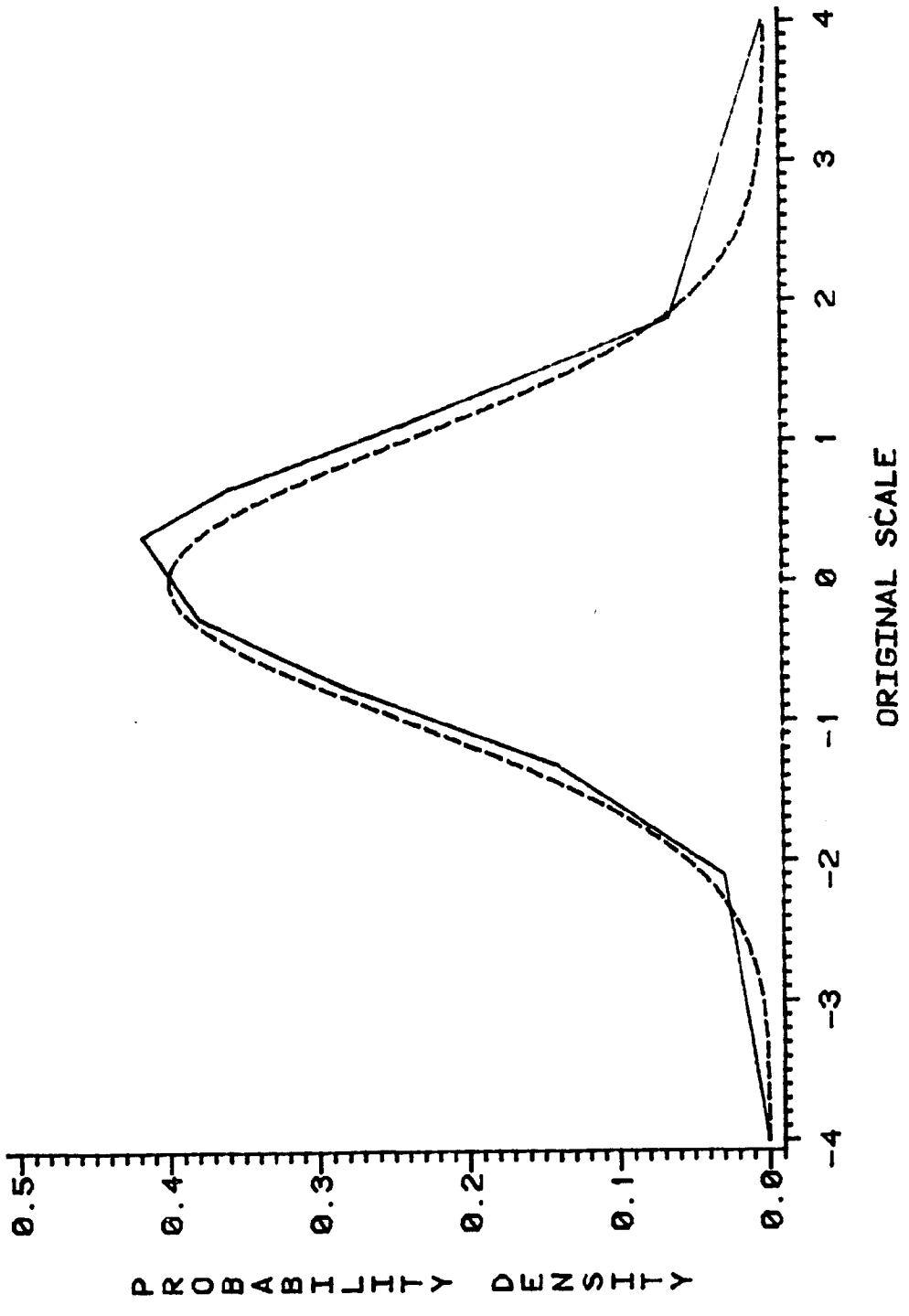

Figure 3.16 Spline density estimation: the spline-estimated density function obtained from a generated sample of 1000 standard-normal deviates. The estimated density is shown by the solid line, and the standard-normal density function is indicated by the broken line.

from them compute the spline distribution functions, $\hat{F}_1(x)$ and $\hat{F}_2(x)$, and the corresponding spline densities, $\hat{f}_1(x)$ and $\hat{f}_2(x)$. Because $\hat{f}_1(x)$ and $\hat{f}_2(x)$ are piece-wise linear functions, the intersection points and $\min\left[\hat{f}_1(x),\hat{f}_2(x)\right]$ can be determined easily, and $\hat{F}_1(x)$ and $\hat{F}_2(x)$ permit ready evaluation of an estimate of OVL, using these points of intersection and the definition of OVL:

$$\tilde{OVL} = \int_x \min\left[\hat{f}_1(x),\hat{f}_2(x)\right]dx \ . \tag{3.17}$$

If the transformations used to map the two sets of sample observations to the interval $[0,1]$ are identical, that is, $g_1(x) = g_2(x) = g(x)$, then the invariance property of OVL can be exploited in the calculation of $\tilde{OVL}$:

$$\tilde{OVL} = \int_{g(x)} \min\left[\hat{f}_1(g(x)),\hat{f}_2(g(x))\right]dg(x) \ . \tag{3.18}$$

The quantity $\min\left[\hat{f}_1(g(x)),\hat{f}_2(g(x))\right]$ can be determined most easily by computing $\hat{f}_1(g(x))$ and $\hat{f}_2(g(x))$ at each point in the sorted union-set of the two sets of breakpoints used in the computation of $\hat{F}_1(g(x))$ and $\hat{F}_2(g(x))$, since a change in the relative positions of $\hat{f}_1(g(x))$ and

$\hat{f}_2(g(x))$ between any two such points necessarily requires that the two estimated densities intersect in the interval so defined. The linear character of $\hat{f}_1(g(x))$ and $\hat{f}_2(g(x))$ makes the determination of this point of intersection trivial.

Consider, for example, the two spline-estimated densities in figure 3.17, which are obtained from two samples of size 100 generated from two normal distributions. The first sample is generated from the standard-normal distribution, and it is the sample used in figures 3.1 through 3.12; the density estimate derived from this sample is indicated by the solid line in figure 3.17. The second sample is generated from a normal distribution with mean 1.0 and variance 4.0; the density estimated from this sample is shown in figure 3.17 by the broken line. It is apparent in this figure that there are two points at which the estimated densities cross. Using the union set of breakpoints and the routine BVALUE, we find these points are 0.251679 and 0.777084 on the transformed scale (equation 3.16 was used to transform both samples). Thus

$$\tilde{OVL} = \hat{F}_1(0.251679) + \hat{F}_2(0.777084) - \hat{F}_2(0.251679) + \hat{F}_1(1.000000) - \hat{F}_1(0.777084) = 0.583812 .$$

The actual overlap between the two normal distributions was calculated as an example in Chapter Two: OVL = 0.609934.

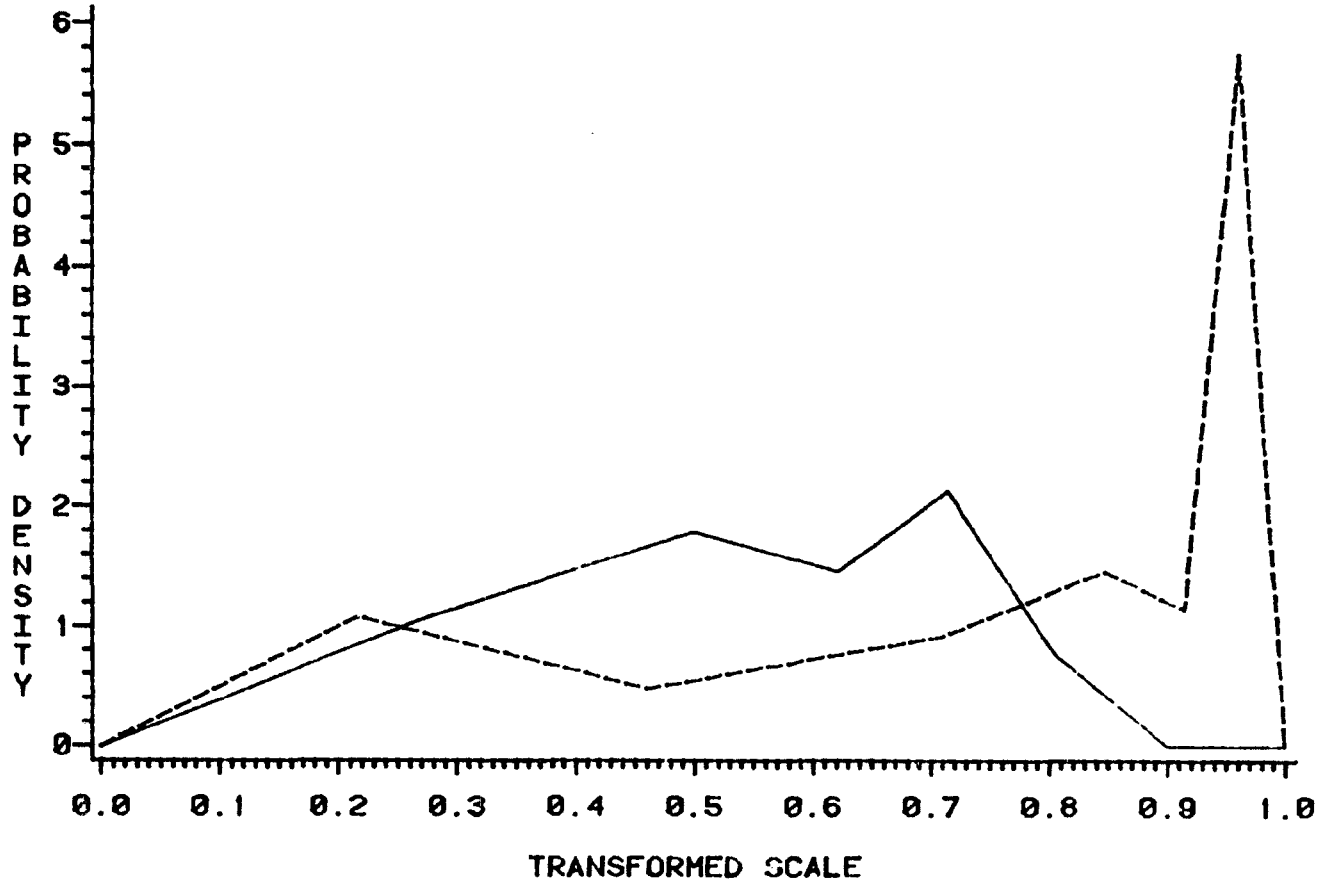

Figure 3.17 Spline estimation of OVL. Two spline-estimated densities are shown, both obtained from generated samples of 100 normal deviates transformed by equation 3.16. The solid line depicts the estimated density, illustrated in figure 3.8, for a sample from the standard-normal distribution, while the broken line indicates the estimated density for a sample from a normal distribution with mean $\mu = 1.0$ and variance $\sigma^2 = 4.0$.

An Estimate of the Variance of $\tilde{OVL}$

Considering the computation complexities inherent to the calculation of $\tilde{OVL}$ from sample data, the problem of determining the sampling variance of $\tilde{OVL}$ is obviously not a simple one. Variance formulae for spline-estimated densities at a specific point based on equally-spaced knot sequences (Lii and Rosenblatt, 1975; Rosenblatt, 1977) or for spline-interpolated densities based on equally-spaced knots (Wahba, 1975) do not apply, and the problem posed by $\tilde{OVL}$--the sum of spline-estimated distribution functions evaluated at points determined from the intersection of their derivatives--quickly suggests that the variance of $\tilde{OVL}$ be estimated indirectly. The method described here, and illustrated in the example below, is based on Efron's development of nonparametric variance estimation procedures (Efron, 1979, 1981, 1982, 1983; Efron and Gong, 1983). Indeed, once the computational set-up for calculating $\tilde{OVL}$ has been realized, the process of generating additional $\tilde{OVL}^*_i$, i=1,...,B, calculated from resamplings of the original sample data and the computation of a bootstrap estimate of the variance of $\tilde{OVL}$, $\hat{Var}_B(\tilde{OVL})$, involves little additional work. It may, however, involve considerable expense, given the computer-intensive calculation of $\tilde{OVL}$.

The idea behind the bootstrap variance estimator is quite simple. We are given the two independent samples, $x_{11},\ldots,x_{1n_1}$ and $x_{21},\ldots,x_{2n_2}$. From these sample data we calculate $\tilde{OVL}$. Now we treat the samples as two finite populations of size n_1 and n_2 respectively, and draw two new bootstrap samples, one from each original sample, <u>with replacement</u>. The sizes of these bootstrap samples are n_1 and n_2 , or the sizes of the

two original samples, giving us the pseudodata, $x^*_{11},\ldots,x^*_{1n_1}$ and $x^*_{21},\ldots,x^*_{2n_2}$. Using this pseudodata, we then calculate the value of our statistic, $\tilde{OVL}$. This resampling procedure is repeated some large number, B, of times; each time we draw a new pair of bootstrap samples from the original data and compute $\tilde{OVL}$. Let $\tilde{OVL}^*_i$ denote the value of $\tilde{OVL}$ computed on the i^{th} iteration of this process. The bootstrap estimator of the variance of $\tilde{OVL}$ is then given by the usual formula for the sample variance:

$$\hat{Var}_B(\tilde{OVL}) = \frac{\sum_{i=1}^{B} (\tilde{OVL}^*_i - \overline{\tilde{OVL}}^*)^2}{B - 1} , \tag{3.19}$$

where

$$\overline{\tilde{OVL}}^* = \frac{\sum_{i=1}^{B} \tilde{OVL}^*_i}{B} . \tag{3.20}$$

The only difficulty with this is, of course, the value of B. As Efron (1982, p. 28) notes, how large "large enough" is depends on and varies from problem to problem, but improvement of the bootstrap

variance estimator is often not great for numbers of bootstrap replications larger than B = 100. The bootstrap method also enables us to do more than simply estimating the variance of $\tilde{OVL}$, and Efron (1982, chaps. 5 and 10) describes how one may use the bootstrap to investigate bias and construct confidence intervals. For these purposes, a value of B much larger than 100 may well prove necessary.

Monte Carlo Investigation of the Properties of $\tilde{OVL}$

To get some idea of the properties of the spline estimator of OVL, $\tilde{OVL}$ has been calculated on a set of Monte Carlo samples from two normal distributions, using a selected number of the design points and sample sizes of the simulation study described in the previous chapter. The design points chosen consist of the four corner points of the original simulation: $\mu_2 = 0.0$, $\sigma_2^2 = 1.0$; $\mu_2 = 1.0$, $\sigma_2^2 = 1.0$; $\mu_2 = 0.0$, $\sigma_2^2 = 3.0$; and $\mu_2 = 1.0$, $\sigma_2^2 = 3.0$. The sample sizes used to investigate the sampling behavior of $\tilde{OVL}$ are $n_1 = n_2 = 100$ and $n_1 = n_2 = 500$. This simulation study permits preliminary assessment of $\tilde{OVL}$ as an estimator of OVL when the two distributions sampled are identical, when the two distributions sampled differ by a substantial difference in their means, when the two distributions sampled differ only by a substantial difference in their variances, and when both the means and variances of the two distributions sampled differ.

On each of the 1000 Monte Carlo trials at each design point and sample size, $\tilde{OVL}$ is computed as described above, using the transformation in (3.16) to map the two independently generated Monte Carlo samples onto the interval $[0,1]$, and calculating $\tilde{OVL}$ on this transformed

scale. The results of this Monte Carlo study are summarized in table 3.1. Comparison of the Monte Carlo means to OVL demonstrates that $\tilde{OVL}$, like $\hat{OVL}$, is a biased estimator of OVL between two normal distributions. Like $\hat{OVL}$, $\tilde{OVL}$ appears to understate OVL, since in only one instance ($\mu_2 = 1.00$, $\sigma_2^2 = 3.0$, $n_1 = n_2 = 500$) does the Monte Carlo mean of $\tilde{OVL}$ exceed OVL. The Monte Carlo variance of $\tilde{OVL}$ presented in table 3.1 is computed from the first two Monte Carlo moments. The variance of $\tilde{OVL}$ decreases as sample sizes increase from $n_1 = n_2 = 100$ to $n_1 = n_2 = 500$, suggesting that $\tilde{OVL}$ is a consistent estimator of OVL. (Because $F_n(x)$ is a consistent estimator of $F(x)$, we should expect $\tilde{OVL}$ to be consistent.)

As before, the bias of $\tilde{OVL}$ is addressed in table 3.1 by computing the standardized bias. Here the standardized bias of $\tilde{OVL}$ is calculated two ways: the difference of the Monte Carlo mean minus OVL divided by the Monte Carlo standard error of $\tilde{OVL}$ (standardized bias, column 1), and this difference divided by the Monte Carlo standard error of $\hat{OVL}$ (standardized bias, column 2). The bias of $\tilde{OVL}$ can be compared to that of $\hat{OVL}$, using the second of these quantities and the standardized bias of $\hat{OVL}$ reproduced in table 3.1 (standardized bias, normal). In units of the standard error of $\hat{OVL}$, then, we see that the bias of $\tilde{OVL}$ is always materially greater than the bias of $\hat{OVL}$, with one exception ($\mu_2 = 1.00$, $\sigma_2^2 = 3.0$, $n_1 = n_2 = 100$).

The relative inefficiency of $\tilde{OVL}$ compared to $\hat{OVL}$ as estimators of OVL between two normal distributions is indicated by the ratio of their Monte Carlo variances, also shown in table 3.1. Evidently the variance of $\tilde{OVL}$ is about 1.5 times the variance of $\hat{OVL}$, running from a low of

TABLE 3.1

RESULTS OF MONTE CARLO SIMULATION STUDY: SPLINE-DENSITY
ESTIMATOR OF OVL BASED ON INDEPENDENT SAMPLES FROM TWO NORMAL DISTRIBUTIONS

$N_1=N_2$	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED BIAS 1	STANDARDIZED BIAS 2	STANDARDIZED BIAS NORMAL	VARIANCE RATIO	KOLMOGOROV STATISTIC
			$\sigma_2^2=1.0$, $\mu_2=0.00$, OVL=1.000000				
100	0.872933	0.00160924	-3.16753	-3.80228	-1.36217	1.440944	0.045761***
500	0.931227	0.00033667	-3.74813	-4.52685	-1.28916	1.458689	0.027065*
			$\sigma_2^2=1.0$, $\mu_2=1.00$, OVL=0.617075				
100	0.605273	0.00330438	-0.20531	-0.22112	-0.02825	1.159925	0.018478
500	0.614617	0.00074031	-0.09034	-0.10364	0.01905	1.315973	0.022620
			$\sigma_2^2=3.0$, $\mu_2=0.00$, OVL=0.740639				
100	0.693836	0.00317019	-0.83125	-1.08225	-0.13483	1.695089	0.024248
500	0.731395	0.00061079	-0.37403	-0.46366	-0.05111	1.536716	0.017141

TABLE 3.1 (CONTINUED)

$N_1=N_2$	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED BIAS 1	STANDARDIZED BIAS 2	STANDARDIZED BIAS NORMAL	VARIANCE RATIO	KOLMOGOROV STATISTIC
			$\sigma_2^2=3.0$, $\mu_2=1.00$, OVL=0.639429				
100	0.634666	0.00291105	-0.08829	-0.10496	-0.12322	1.413392	0.022382
500	0.642406	0.00073362	0.10991	0.14102	-0.03012	1.646285	0.013137

NOTE: ASTERISKS DENOTE REJECTION OF THE NORMAL DISTRIBUTION AT THE 0.10 (*), 0.05 (**), AND 0.01 (***) LEVELS OF SIGNIFICANCE USING THE MODIFIED KOLMOGOROV STATISTIC AND THE PSEUDOCRITICAL VALUES IN STEPHANS (1974) FOR NORMALITY, MEAN AND VARIANCE UNKNOWN.

1.15 ($\mu_2 = 1.00$, $\sigma_2^2 = 1.0$, $n_1 = n_2 = 100$) to a high of 1.70 ($\mu_2 = 0.00$, $\sigma_2^2 = 3.0$, $n_1 = n_2 = 100$). With only four points in this simulation study, we cannot say much about the relative efficiency of $\tilde{OVL}$ versus $\hat{OVL}$ as a function of the difference in population means, the difference in population variances, and sample size, except to note that the ratio of the Monte Carlo variances of $\tilde{OVL}$ and $\hat{OVL}$ changes little with the increase in sample sizes from $n_1 = n_2 = 100$ to $n_1 = n_2 = 500$ in the case of sampling from identical normal distributions, increases with sample size when sampling from normal distributions with different means, and decreases as sample sizes increase when sampling from normal distributions with the same mean but different variances. Finally, as the Kolmogorov statistics in table 3.1 indicate, $\tilde{OVL}$ exhibits, at least approximately, a normal sampling distribution when sampling from sufficiently dissimilar normal distributions. Normality does not hold when the two distributions sampled are the same and sample sizes are small, but the normality or nonnormality of $\tilde{OVL}$ when small differences in means or variances distinguish the distributions from which the two samples arise obviously cannot be determined from this Monte Carlo simulation study.

Discussion

The results of the Monte Carlo investigation of the behavior of $\tilde{OVL}$ suggest that the spline estimator of OVL can perform well. The properties of $\tilde{OVL}$ appear to echo those displayed by $\hat{OVL}$; in particular, the bias of $\tilde{OVL}$ is related to OVL and the sample sizes in the same way as the bias of $\hat{OVL}$. As expected, the variance of $\tilde{OVL}$ exceeds the variance of $\hat{OVL}$ when sampling from two normal distributions. Since

the primary advantage of $\tilde{OVL}$ over $\hat{OVL}$ is its distribution-free approach, the performance of $\tilde{OVL}$ relative to $\hat{OVL}$ in the normal case indicates that $\tilde{OVL}$ should perform adequately in situations of more immediate interest, where $\hat{OVL}$ is an inappropriate estimator of OVL. As the example below demonstrates, the spline-density estimator of OVL, combined with the bootstrap technique of estimating its variance and constructing confidence intervals, can indeed prove worthwhile in real problems of data analysis. The bias-corrected percentile method of constructing confidence intervals for the true overlap between the unknown distributions from the bootstrap distribution of $\tilde{OVL}$ may, in fact, counter-balance the apparent increase in the downward bias of the estimator of OVL when nonparametric estimation is adopted.

The success of the spline-density based technique of estimating OVL raises the possibility that a less sophistocated nonparamteric method might also prove adequate in problem settings where distributional assumptions seem unwarranted. An obvious alternative to the spline estimator of the unknown densities is the kernal method of density estimation. A number of kernal functions can be used in this latter approach, but the "naive" kernal estimator offers a simple, and perhaps entirely adequate, technique for estimating the two densities required for the computation of an estimate of OVL (Rosenblatt, 1956; Waterman and Whiteman, 1978). If such density estimates are used to obtain the points of intersection of the densities, then the sample distribution functions themselves could be employed to evaluate the necessary components of the estimated overlapping coefficient, and the bootstrap can again be used to estimate the variance of the estimator of OVL and to construct confidence regions for the unknown OVL.

For the present, the potential of these alternative nonparametric approaches for the estimation of OVL must remain an open question.

An Example

Let us now reconsider the wealth example introduced in Chapter Two. The reader will recall that the estimate of OVL, assuming normality and equal population variances, was $\hat{OVL} = 0.859614$. Here we shall compute $\tilde{OVL}$ from these data and use Efron's bootstrap methodology to obtain an estimate of the standard error of $\tilde{OVL}$ and to construct confidence intervals for the unknown overlap between the distributions of wealth of the persistent and nonpersistent Alabama farmers.

The transformation chosen to map the natural logarithms of estimated 1850 wealth onto the interval $[0,1]$ is that given in equation 3.14, with $a = 3$ and $b = 12$. (Note that the smallest observation in the combined samples is 3.22865 and the largest observed natural logarithm of wealth is 11.47963, both in the nonpersistent group.) Thus the data actually used to compute $\tilde{OVL}$ were obtained from the following transformation:

$$g(x) = \frac{\log_e(x) - 3}{9} \quad .$$

This transformation is applied to the wealth data for both persistent and nonpersistent farmers, and $\tilde{OVL}$ calculated as described above. From these data, $\tilde{OVL} = 0.869152$.

Bootstrap estimates of the variance of $\tilde{OVL}$ are readily obtained by equation 3.19. Here the bootstrap resamplings are accomplished with the simple FORTRAN subroutine RESAMP in the Appendix. Results for three different values of B are the following:

$$B = 100,\ \overline{\tilde{OVL}^*} = 0.843070,\ \text{and}\ \hat{Var}_B(\tilde{OVL}) = 0.00112578;$$

$$B = 250,\ \overline{\tilde{OVL}^*} = 0.842914,\ \text{and}\ \hat{Var}_B(\tilde{OVL}) = 0.000961849;$$

$$B = 500,\ \overline{\tilde{OVL}^*} = 0.842576,\ \text{and}\ \hat{Var}_B(\tilde{OVL}) = 0.000966696.$$

This example demonstrates that the bootstrap estimate of the variance of $\tilde{OVL}$ is fairly good (that is, close to what is obtained for larger B) when B = 100, but that a larger value of B is preferable. If we use the result obtained when B = 500, the estimated standard error of $\tilde{OVL}$ is 0.0310917. To provide some indication of the the cost of finding this bootstrap estimate of the standard error of $\tilde{OVL}$, the computation of $\tilde{OVL}$ and the generation of 500 $\tilde{OVL}^*$ with the FORTRAN routines in the Appendix required slightly more than seven minutes of CPU time on an IBM 4381-2 for this example.

Two of the methods described by Efron (1982, chap. 10) will be used to construct bootstrap confidence intervals for OVL using the 1850 wealth data. Let $F_B^*(\cdot)$ be the empirical distribution function constructed from the $\tilde{OVL}_i^*$ (i=1,...,B), and let $F_B^{*-1}(\cdot)$ denote its inverse. A $(1 - \alpha)100\%$ confidence interval for OVL using the percentile method is

$$\left[F_B^{*-1}(\alpha/2), F_B^{*-1}(1 - \alpha/2)\right] . \qquad (3.21)$$

Thus a 90% confidence interval for the true overlap between the wealth distributions of the persistent and nonpersistent Alabama farmers, using the bootstrap distribution function constructed from the 500 $\tilde{OVL}^*$, is given by

$$\left[F^{*-1}_{500}(0.05), F^{*-1}_{500}(0.95)\right] = (0.792479, 0.895659) \ ;$$

see figure 3.18.

The second method of constructing confidence intervals is what Efron calls the bias-corrected percentile method. Let $\Phi^{-1}(\cdot)$ denote the inverse standard-normal distribution function. Define

$$z_0 \equiv \Phi^{-1}(F^*_B(\tilde{OVL})) \ , \tag{3.22}$$

and

$$z_{\alpha/2} \equiv \Phi^{-1}(1 - \alpha/2) \ . \tag{3.23}$$

Then the $(1 - \alpha)100\%$ bias-corrected confidence interval for OVL

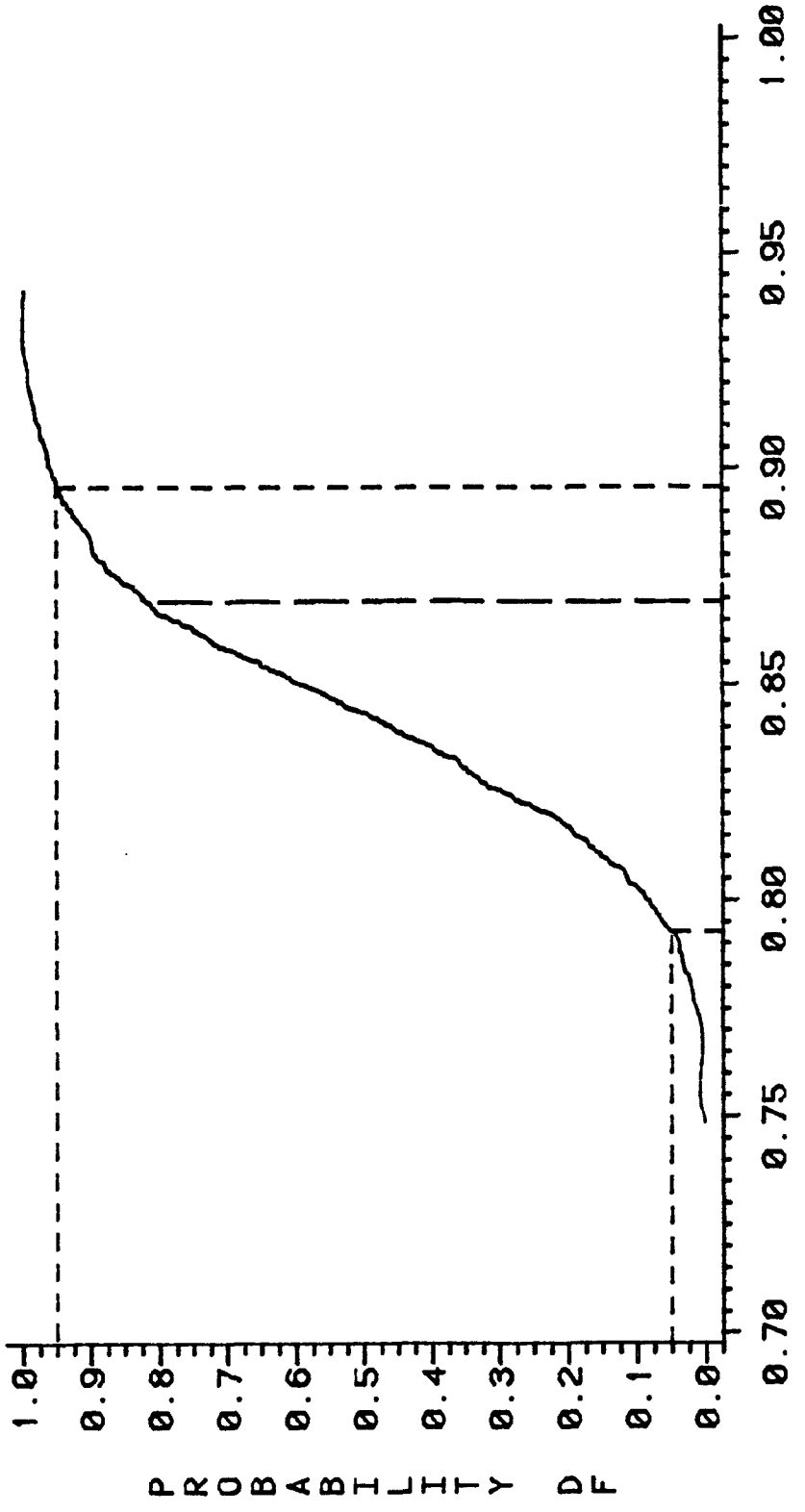

Figure 3.18 Construction of a 90% confidence interval for the overlap between the distributions of wealth for persistent and nonpersistent Alabama farmers in 1850 by the percentile method. The bootstrap distribution function for the spline estimator of OVL obtained from the wealth data (B = 500) is shown by the solid line. The estimated OVL is indicated by the heavy broken line, and the limits of the confidence interval by the lighter broken lines.

given by the following:

$$\left[F_B^{*-1}(\Phi(2z_0 - z_{\alpha/2})), F_B^{*-1}(\Phi(2z_0 - z_{\alpha/2}))\right] . \qquad (3.24)$$

Since $\tilde{OVL}$ computed from the wealth data is 0.869152 and $F^*_{500}(0.869152)$ is 0.818363, here $z_0 = 0.909145$. If we want a 90% confidence interval for OVL, $\alpha = 0.10$ and $z_{\alpha/2} = 1.64485$. The 90% bias-corrected confidence interval for the true overlap between the wealth distributions of the persistent and nonpersistent Alabama farmers is

$$\left[F_{500}^{*-1}(\Phi(0.173440)), F_{500}^{*-1}(\Phi(3.46314))\right] =$$

$$= \left[F_{500}^{*-1}(0.568846), F_{500}^{*-1}(0.999733)\right]$$

$$= (0.848472, 0.941238) ;$$

see figure 3.19.

Note that the 90% confidence interval for OVL obtained by the percentile method is close to the interval obtained in Chapter Two by normal theory; the limits of the percentile confidence interval, however, are slightly below the corresponding normal theory limits of 0.808967 and 0.915465. On the other hand, the limits of the bias-corrected confidence interval for OVL are more distant from the normal

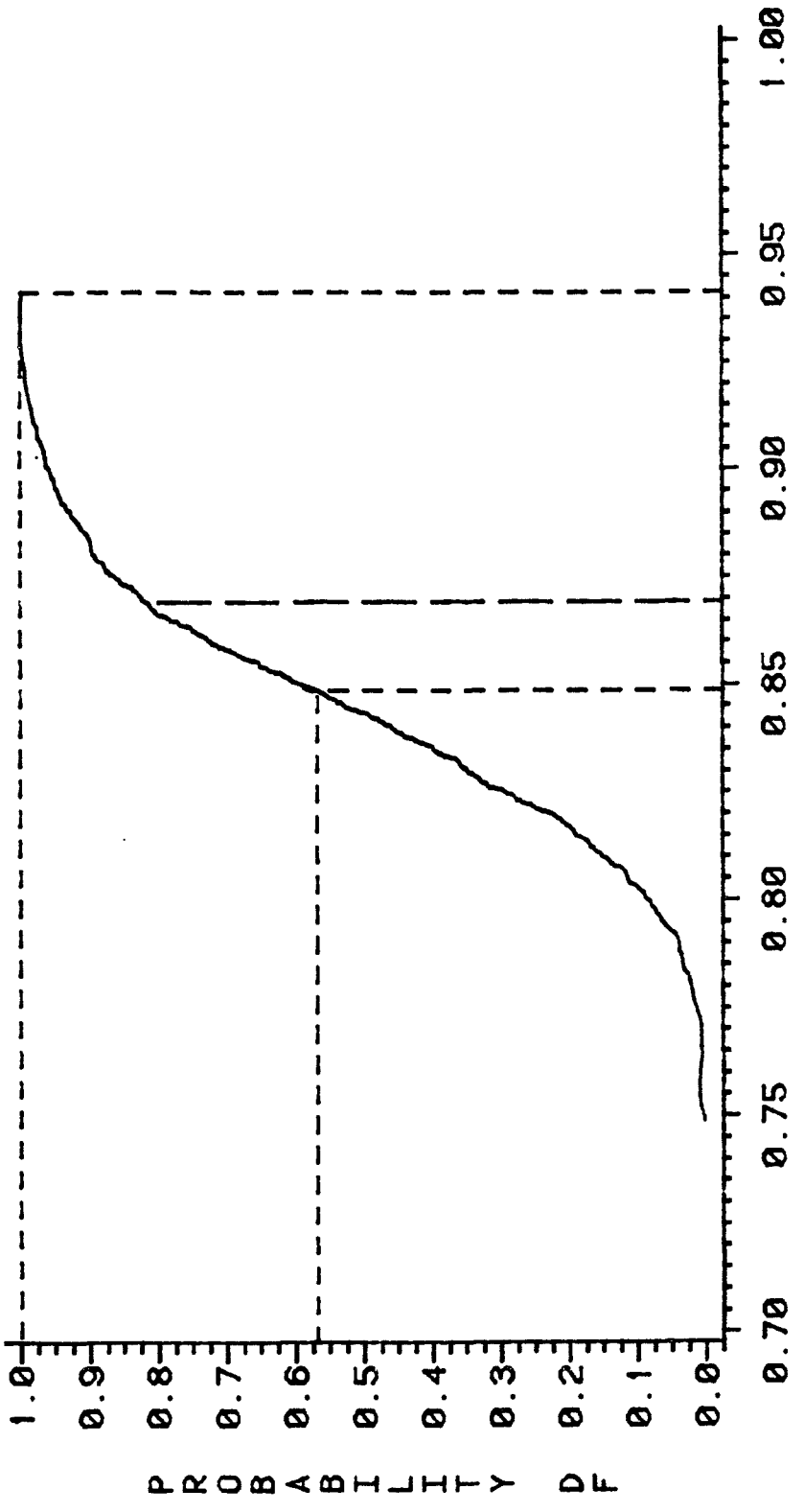

Figure 3.19 Construction of a 90% confidence interval for the overlap between the distributions of wealth for persistent and nonpersistent Alabama farmers in 1850 by the bias-corrected percentile method. The bootstrap distribution function for the spline estimator of OVL obtained from the wealth data ($B = 500$) is shown by the solid line. The estimated OVL is indicated by the heavy broken line, and the limits of the confidence interval by the lighter broken lines.

theory limits, and the bias-corrected percentile method produces an interval with upper and lower endpoints higher than either the normal theory or the percentile limits. Given our current knowledge of the sampling behavior of $\tilde{OVL}$, it is impossible to conclude that one of these confidence intervals is superior to the other two in every problem setting, but the evident nonnormality of these sample data and the general downward bias of estimators of OVL suggest that the bias-corrected confidence interval for the overlap between the distributions of wealth of persistent and nonpersistent Alabama farmers in 1850 is the most realistic of the three 90% confidence intervals constructed from the wealth data.

Chapter Four

OVL AS A MEASURE OF ASSOCIATION IN A 2 X C CONTINGENCY TABLE

The investigation of OVL as a measure of agreement between two distributions arranged in a 2 X C contingency table addresses the behavior of the overlapping coefficient in the context in which it was first proposed and used (Weitzman, 1970). The properties of the estimator of OVL when the table entries are regarded as random variables is examined for two probability models of the 2 X C table, and in each case it is more convenient to use the index of dissimilarity, D (D = 1 - OVL), instead of OVL itself. The two probability models considered are first, when the rows of the table are independent realizations of two possibly identical multinomial distributions (row totals fixed), and second, when the cells of the table are determined by the multivariate hypergeomtric distribution of a single row of the table (both row totals and column totals fixed). Based on the behavior of D in the 2 X C table under these assumptions, it is apparent that in both situations the estimator of OVL is biased, that this bias means the estimator of OVL understates the true overlap of the row distributions, that the magnitude of the bias is directly related to C and the true overlap between the row distributions (unity in the multivariate hypergeometric model) and declines as the sizes of the row totals become

large. Similarly, the variance of the estimator of OVL increases with C and decreases as the row totals increase.

OVL and D as Measures of Association Between Two Categorized Populations

The overlap between two Poisson distributions served as one of the examples introduced earlier as an illustration of the computation of the overlapping coefficient between two known distributions. Here a special case of OVL involving the overlap between two discrete distributions will be examined. Let us begin by supposing that we wish to compare two finite populations whose elements can be classified into C categories. One can think of this situation as the cross-classification of these two populations into a 2 X C contingency table, the two rows of the table representing the two populations and the C columns of the table representing the C categories into which the populations are sorted. Let n_{ij} denote the number of individuals from population (row) i falling into category (column) j of this table. Finally, let N_1 and N_2 denote the population (row) totals for the first and second populations respectively. The overlap between the two populations may then be computed as follows:

$$OVL = \sum_{j=1}^{C} \min\left(\frac{n_{1j}}{N_1}, \frac{n_{2j}}{N_2}\right). \tag{4.1}$$

However, it will be more convenient in the discussion that follows if

we use the relationship between OVL and D, the index of dissimilarity, and work with D rather than OVL. The usual formulation of D in the circumstance just described is

$$D = \frac{1}{2} \sum_{j=1}^{C} \left| \frac{n_{1j}}{N_1} - \frac{n_{2j}}{N_2} \right| . \qquad (4.2)$$

Because OVL = 1 - D, it follows that any results we obtain for one automatically apply to the other.

In the context of the 2 X C table, OVL and D are simply two of many proposed measures of association (Goodman and Kruskal, 1979). They have the advantage that they are relatively easy to compute, and both remain unaffected by row or column permutation or the multiplication of an entire row by some nonzero constant. The measure D also appears to possess a natural meaning for many users: D represents the minimum proportion of individuals in either population (row) whose reclassification into the appropriate categories (columns) would produce two populations with equal proportions in each category (Taeuber and Taeuber, 1976, pp. 887-88; Goodman and Kruskal, 1979, p. 56). The properties of the index of dissimilarity as a measure of association, particularly as an indicator of residential segregation by race, have generated both controversy and confusion in the sociological literature (Jahn et al., 1947, 1948; Hornseth, 1947; Williams, 1948; Jahn, 1950; Duncan and Duncan, 1955; Taeuber and Taeuber, 1965, 1976; Cortese et al., 1976,

1978; Winship, 1977, 1978; Massey, 1978; Falk et al., 1978; Elgie, 1979; Kestenbaum, 1980; Merschrod, 1981). Nevertheless, D has been extended to tables with more than two rows (Morgan and Norbury, 1981; Sakoda, 1981), and it is used in a variety of applications, if not always defined and computed correctly (Hout, 1983, pp. 12-13).

In the following discussion, properties of D--and thus OVL--will be investigated in two circumstances where the cell counts of the 2 X C table are presumed to follow a specified probability law. In the first instance, the two rows of the table are treated as independent realizations of two possibly identical multinomial distributions, with the row totals fixed. The second situation to be examined is the multivariate hypergeometric model proposed by Cortese et al. (1976) for a 2 X C table with both row and column totals fixed. To indicate that the entries of the table will now be treated as random variables, x_{ij} (i=1,2; j=1,...,C) will denote the number of individuals in row i and column j of the 2 X C contingency table.

The Multinomial Model of the 2 X C Table

Let us assume that the two rows of the 2 X C table are independent realizations of two multinomial distributions, possibly identical. Let N_1 and N_2, the two row totals, be fixed. Then the probability law for either row of the table is given by

$$P(x_{i1},\ldots,x_{iC}) = N_i! \prod_{j=1}^{C} \frac{p_{ij}^{x_{ij}}}{x_{ij}!}, \quad 0 \le x_{ij} \le N_i;$$

$$i=1,2;\ j=1,\ldots,C; \tag{4.3}$$

subject to the condition

$$\sum_{j=1}^{C} x_{ij} = N_i \ .$$

It can be shown that (Johnson and Kotz, 1969, pp. 51, 284)

$$E(x_{ij}) = N_i p_{ij} \ , \quad i=1,2;\ j=1,\ldots,C; \tag{4.4}$$

$$Var(x_{ij}) = N_i p_{ij}(1 - p_{ij}) \ , \quad i=1,2;\ j=1,\ldots,C; \tag{4.5}$$

and

$$Cov(x_{ij},x_{ij'}) = - N_i p_{ij} p_{ij'} \ ,$$

$$i=1,2;\ j=1,\ldots,C;\ j'=1,\ldots,C;\ j \neq j' \ . \tag{4.6}$$

Because of the assumed row independence, we note

$$\mathrm{Cov}(x_{1j}, x_{2j'}) = 0 \ , \quad j=1,\ldots,C; \ j'=1,\ldots,C. \tag{4.7}$$

This probability model for the 2 X C table arises naturally when the two rows represent two independent simple random samples of sizes N_1 and N_2 classified into the C categories represented by the columns of the table, with the unknown proportion of each population in each of the C categories given by p_{ij} (i=1,2; j=1,...,C). The true value of D for the two sampled populations, of course, is given by

$$D = \frac{1}{2} \sum_{j=1}^{C} |p_{1j} - p_{2j}| \ . \tag{4.8}$$

The index of dissimilarity may be calculated for any realization of the table generated by (4.3); let us denote this statistic by $\hat{D}$:

$$\hat{D} = \frac{1}{2} \sum_{j=1}^{C} |d_j| \ , \tag{4.9}$$

where

$$d_j = \frac{x_{1j}}{N_1} - \frac{x_{2j}}{N_2}, \quad j=1,\ldots,C. \tag{4.10}$$

In the case of two independent simple random samples classified into the C categories, $\hat{D}$ is the maximum-likelihood estimator of D given in equation 4.8, and x_{ij}/N_i (i=1,2; j=1,...,C) is the maximum-likelihood estimator of p_{ij}.

Expectation and Variance of $\hat{D}$

By definition, the mean and variance of $\hat{D}$ are given by

$$E(\hat{D}) = \frac{1}{2} \sum_{j=1}^{C} \mu_j , \tag{4.11}$$

$$\mathrm{Var}(\hat{D}) = \frac{1}{4}\left[\sum_{j=1}^{C} \sigma_j^2 + \sum_{j\neq j}^{C}\sum \sigma_{jj'}\right] ; \tag{4.12}$$

where μ_j, σ_j^2, and $\sigma_{jj'}$ are defined as follows:

$$\mu_j = E(|d_j|) \ , \quad j=1,\ldots,C; \tag{4.13}$$

$$\sigma_j^2 = Var(|d_j|) \ , \quad j=1,\ldots,C; \tag{4.14}$$

and

$$\sigma_{jj'} = Cov(|d_j|,|d_{j'}|) \ , \quad j=1,\ldots,C; \ j'=1,\ldots,C; \ j\neq j'. \tag{4.15}$$

Now from (4.3), the distribution of any x_{ij} is binomial, with probability function

$$P(x_{ij}) = \binom{N_i}{x_{ij}} p_{ij}^{x_{ij}} \cdot (1 - p_{ij})^{N_i - x_{ij}} \ ,$$

$$0 \leq x_{ij} \leq N_i; \ i=1,2; \ j=1,\ldots,C. \tag{4.16}$$

Furthermore, the joint distribution of any x_{ij} and $x_{ij'}$ $(j \neq j')$ is multinomial, with probability function

$$P(x_{ij},x_{ij'}) = \frac{N_i! \cdot p_{ij}^{x_{ij}} \cdot p_{ij'}^{x_{ij'}} \cdot (1 - p_{ij} - p_{ij'})^{N_i - x_{ij} - x_{ij'}}}{x_{ij}! \cdot x_{ij'}! \cdot (N_i - x_{ij} - x_{ij'})!},$$

$$0 \le x_{ij} + x_{ij'} \le N_i;\ i=1,2;$$

$$j=1,\ldots,C;\ j'=1,\ldots,C;\ j \neq j'. \qquad (4.17)$$

Because the rows are assumed to be independent, the joint distribution of x_{1j} and x_{2j} is simply the product of the marginal probabilities given by equation 4.16. Therefore the expectation of $|d_j|$ is

$$\mu_j = \sum_{x_{1j}=0}^{N_1} \sum_{x_{2j}=0}^{N_2} \left| \frac{x_{1j}}{N_1} - \frac{x_{2j}}{N_2} \right| \cdot P(x_{1j}) \cdot P(x_{2j}) \ . \qquad (4.18)$$

Because $E(|d_j|^2) = E(d_j^2)$, the variance of $|d_j|$ is given by

$$\sigma_j^2 = \mathrm{Var}(d_j) + \left[E(d_j)\right]^2 - \left[E(|d_j|)\right]^2 \ . \qquad (4.19)$$

From (4.10) and the assumption of row independence, it follows that

$$\sigma_j^2 = \frac{p_{1j}(1 - p_{1j})}{N_1} + \frac{p_{2j}(1 - p_{2j})}{N_2} + (p_{1j} - p_{2j})^2 - \mu_j^2 . \qquad (4.20)$$

Finally, because $|d_j| \cdot |d_{j'}| = |d_j d_{j'}|$, the covariance of $|d_j|$ and $|d_{j'}|$ $(j \neq j')$ can be written as

$$\sigma_{jj'} = \sum_{x_{1j}=0}^{x_{1j}+x_{1j'} \leq N_1} \sum_{x_{1j'}=0} \sum_{x_{2j}=0}^{x_{2j}+x_{2j'} \leq N_2} \sum_{x_{2j'}=0} \left[\left| \frac{x_{1j}x_{1j'}}{N_1^2} - \frac{x_{1j}x_{2j'} + x_{1j'}x_{2j}}{N_1 N_2} + \frac{x_{2j}x_{2j'}}{N_2^2} \right| \cdot P(x_{1j}, x_{1j'}) \cdot P(x_{2j}, x_{2j'}) \right] - \mu_j \mu_{j'} . \qquad (4.21)$$

Substitution of equations 4.18, 4.20, and 4.21 into equations 4.11 and 4.12 yields the expectation and variance of $\hat{D}$.

Normal Approximation to the Mean and Variance of $\hat{D}$

The expressions for the expectation and variance of $\hat{D}$ derived above require extensive computation for nontrivial C, N_1, and N_2. Equation 4.21 in particular proves difficult, as there are $C(C - 1)/2$ unique covariance terms $\sigma_{jj'}$ to be calculated, each involving quadruple summation over N_1 and N_2. Considerably simpler expressions for the mean of $\hat{D}$ and, in a special case of some interest, for the variance of $\hat{D}$ can be obtained by using a multivariate-normal approximation to the

multinomial row distributions given by (4.3). That is, we assume the distribution of $x_{i1},\ldots,x_{iC}$ (i=1,2) is C-variate normal with elements of the mean vector given by equation 4.4 and elements of the variance-covariance matrix given by equations 4.5 and 4.6. This immediately implies that $d_1,\ldots,d_C$ are distributed as C-variate normal and that $|d_1|,\ldots,|d_C|$ are distributed as C-variate folded-normal.

We argue as follows. Each $d_j = \frac{x_{1j}}{N_1} - \frac{x_{2j}}{N_2}$ will be (approximately) normally distributed with expectation ξ_j,

$$\xi_j = p_{1j} - p_{2j} , \tag{4.22}$$

and variance τ_j^2,

$$\tau_j^2 = \frac{p_{1j}(1 - p_{1j})}{N_1} + \frac{p_{2j}(1 - p_{2j})}{N_2} . \tag{4.23}$$

It follows from our assumption of the approximate C-variate normality of $x_{11},\ldots, x_{1C}$ and $x_{21},\ldots,x_{2C}$ that the d_j are normally distributed and thus that the $|d_j|$ are distributed as the folded normal distribution. The mean μ_j of each $|d_j|$ is obtained from ξ_j and τ_j^2 by equation 2.44. Therefore the approximate expectation of $\hat{D}$ is provided by

$$E(\hat{D}) \doteq \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{C} \left\{ \tau_j \cdot \exp\left(\frac{-\xi_j^2}{2\tau_j^2}\right) + \xi_j \left[1 - 2\Phi\left(\frac{-\xi_j}{\tau_j}\right)\right]\right\} . \qquad (4.24)$$

If $D = 0$, that is if $p_{1j} = p_{2j} = p_j$ $(j=1,\ldots,C)$, then

$$E(\hat{D}) \doteq \left(\frac{N_1 + N_2}{2\pi N_1 N_2}\right)^{\frac{1}{2}} \sum_{j=1}^{C} [p_j(1 - p_j)]^{\frac{1}{2}} . \qquad (4.25)$$

This latter expression is always positive, except in the pathological case when one $p_j = 1$ and all others are equal to zero, and therefore $\hat{D}$ will always be biased above as an estimator of D when the two distributions from which the two rows are obtained are identical. Of course, this bias of $\hat{D}$ can be made to decrease by increasing N_1 and N_2. Because the elements of the sum in equations 4.24 and 4.25 are positive, the bias of $\hat{D}$ is directly related to C, the number of columns in the table, and to the p_{ij} $(i=1,2;\ j=1,\ldots,C)$, the multinomial probabilities.

The properties of the folded-normal distribution, discussed in Chapter Two, suggest that the behavior of $\hat{D}$ when $D = 0$ represents the extreme case in regard to bias. Thus as the magnitude of ξ_j/τ_j increases, μ_j will approach $|\xi_j| = |p_{1j} - p_{2j}|$, and so the expected value of $\hat{D}$ will approach D as the magnitudes of all ξ_j/τ_j $(j=1,\ldots,C)$ become large. Once again, then, the bias of $\hat{D}$ is apparently least when the two distributions compared are sufficiently dissimilar, the sample

sizes (row totals) N_1 and N_2 are sufficiently large, or some combination of the two, provided, of course, that $p_{1j} \neq p_{2j}$ (j=1,...,C). It is also evident that $\hat{D}$ may exhibit approximate normality if $\mu_j/\tau_j > 3$ for all j (j=1,...,C), since each $|d_j|$ becomes approximately normally distributed whenever this may hold, given our assumption about the normality of the $x_{11},\ldots,x_{1C}$ and $x_{21},\ldots,x_{2C}$ (Elandt, 1961).

An approximation to the variance of $\hat{D}$ is also possible when D = 0. In this instance, $\xi_j = 0$ (j=1,...,C), and equation 4.23 can be written as

$$\tau_j^2 = \frac{N_1 + N_2}{N_1 N_2} p_j(1 - p_j) \quad . \tag{4.26}$$

Because of the presumed row independence, the covariance of d_j and $d_{j'}$ ($j \neq j'$) follows from (4.6) and (4.7):

$$\tau_{jj'} = \mathrm{Cov}(d_j, d_{j'}) = - \frac{N_1 + N_2}{N_1 N_2} p_j p_{j'} \quad . \tag{4.27}$$

From equation 4.26 and equation 4.27, $\rho_{jj'}$, the correlation between d_j and $d_{j'}$ ($j \neq j'$), is

$$\rho_{jj'} = -\left[\frac{p_j p_{j'}}{(1 - p_j)(1 - p_{j'})}\right]^{\frac{1}{2}} . \tag{4.28}$$

From equations 2.44 and 2.45, the approximate variance of $|d_j|$ is given by

$$\sigma_j^2 \doteq \frac{N_1 + N_2}{N_1 N_2}\left(\frac{\pi - 2}{\pi}\right) p_j(1 - p_j) , \tag{4.29}$$

and from the absolute moments of the multivariate-normal distribution with zero mean vector (Nabeya, 1951, 1952), the approximate covariance of $|d_j|$ and $|d_{j'}|$ $(j \neq j')$ is given by

$$\begin{aligned}\sigma_{jj'} &\doteq \frac{2}{\pi}\left[(1 - \rho_{jj'}^2)^{\frac{1}{2}} + \rho_{jj'}\sin^{-1}(\rho_{jj'})\right]\tau_j\tau_{j'} - \mu_j\mu_{j'} \\ &= \frac{2(N_1 + N_2)}{\pi N_1 N_2}\left[p_j(1 - p_j)p_{j'}(1 - p_{j'})\right]^{\frac{1}{2}}\cdot\left[(1 - \rho_{jj'})^{\frac{1}{2}} + \right. \\ &\qquad \left. + \rho_{jj'}\sin^{-1}(\rho_{jj'}) - 1\right] .\end{aligned} \tag{4.30}$$

Substitution of (4.29) and (4.30) into equation 4.12 yields the following expression for the variance of $\hat{D}$ when $D = 0$:

$$\mathrm{Var}(\hat{D}) \doteq \frac{N_1 + N_2}{4\pi N_1 N_2}\left\{(\pi - 2)\left(1 - \sum_{j=1}^{C} p_j^2\right) + \right.$$

$$+ 2\sum_{j\neq j'}^{C}\sum \left[p_j(1 - p_j)p_{j'}(1 - p_{j'})\right]^{\frac{1}{2}}\cdot\left[(1 - \rho_{jj'}^2)^{\frac{1}{2}} + \right.$$

$$\left.\left. + \rho_{jj'}\sin^{-1}(\rho_{jj'}) - 1\right]\right\}. \qquad (4.31)$$

Inspection of this expression for the approximate variance of $\hat{D}$ reveals that $\mathrm{Var}(\hat{D})$ is directly related to the common multinomial probabilities, p_j $(j=1,\ldots,C)$; is an increasing function of C; and decreases as N_1 and N_2 increase.

Monte Carlo Investigation of Properties of $\hat{D}$

To examine the sampling behavior of $\hat{D}$ (and thus of $\hat{OVL}$ defined in the 2 X C table with independent multinomial row distributions), a Monte Carlo simulation study involving 1000 Monte Carlo trials at each of 48 design points was undertaken. The objectives of this study are to determine how the bias and sampling variance of $\hat{D}$ vary with D, C, and the sample sizes (row totals) N_1 and N_2; to assess the utility of the normal approximation to the expectation of $\hat{D}$; and to investigate the possibility of approximating the sampling distribution of $\hat{D}$ with some appropriately specified continuous probability law.

Four different values of D, 0.05, 0.25, 0.45, and 0.65, and three values of C, 4, 7, and 11, were chosen so that the behavior of $\hat{D}$ could be evaluated when the real association between the row distributions is

high, moderate, and low; and where the number of columns in the table remains small enough to permit simple assessment of the effect of the size of the table (C = 4 and C = 7) and large enough to represent a more realistic setting for the actual use of $\hat{D}$ (C = 11). Multinomial probabilities for each row of the table were then fixed to obtain the desired D at every value of C; see table 4.1. At each combination of D and C, four sets of values for N_1 and N_2 are used to generate the Monte Carlo distributions of D: $N_1 = N_2 = 100$; $N_1 = 100$, $N_2 = 200$; $N_1 = 200$, $N_2 = 100$; and $N_1 = N_2 = 200$. (Note that when the multinomial probabilities assigned to the two rows are reverse images of each other, as for D = 0.05 and C = 4, the sets of Monte Carlo trials with unequal sample sizes are actually replications of the same design point.) All sets of Monte Carlo trials were generated using the MATRIX procedure in SAS (SAS, 1982).

The results of the Monte Carlo simulation study are summarized in table 4.2. The Monte Carlo mean and variance of $\hat{D}$ are based on the first and second moments computed from the 1000 realizations of $\hat{D}$ at each combination of D, C, N_1, and N_2. Direct inspection of the Monte Carlo means in table 4.2 demonstrates that $\hat{D}$ is a biased estimator of D, just as equation 4.24 suggests. The bias of $\hat{D}$, nearly without exception, declines as N_1 and N_2 increase, but it remains substantial when D = 0.05. The bias of $\hat{D}$ measured in units of the Monte Carlo standard error of $\hat{D}$, the standardized bias in table 4.2, indicates not only that the bias of $\hat{D}$ declines absolutely as N_1 and N_2 increase, but also that this bias declines relative to the sampling error of $\hat{D}$. The decline in both the bias and the standardized bias is sometimes erratic,

TABLE 4.1

MULTINOMIAL PROBABILITIES USED IN THE MONTE CARLO SIMULATION STUDY
OF THE INDEX OF DISSIMILARITY IN A 2 X C TABLE WITH INDEPENDENT MULTINOMIAL ROW DISTRIBUTIONS

C	ROW	P_{i1}	P_{i2}	P_{i3}	P_{i4}	P_{i5}	P_{i6}	P_{i7}	P_{i8}	P_{i9}	P_{i10}	P_{i11}
						D = 0.05						
4	1	0.200	0.275	0.300	0.225	.	.	.	.	.	.	.
	2	0.225	0.300	0.275	0.200	.	.	.	.	.	.	.
7	1	0.100	0.150	0.200	0.175	0.135	0.130	0.110	.	.	.	.
	2	0.110	0.165	0.225	0.150	0.125	0.125	0.100	.	.	.	.
11	1	0.025	0.050	0.075	0.125	0.200	0.185	0.130	0.090	0.060	0.030	0.030
	2	0.030	0.060	0.090	0.130	0.215	0.175	0.125	0.075	0.050	0.025	0.025
						D = 0.25						
4	1	0.125	0.250	0.325	0.300	.	.	.	.	.	.	.
	2	0.300	0.325	0.250	0.125	.	.	.	.	.	.	.
7	1	0.100	0.125	0.150	0.250	0.150	0.125	0.100	.	.	.	.
	2	0.175	0.225	0.225	0.125	0.100	0.080	0.070	.	.	.	.
11	1	0.025	0.050	0.075	0.100	0.125	0.125	0.150	0.125	0.100	0.075	0.050
	2	0.050	0.100	0.125	0.200	0.150	0.100	0.085	0.070	0.055	0.040	0.025

TABLE 4.1 (CONTINUED)

C	ROW	P_{i1}	P_{i2}	P_{i3}	P_{i4}	P_{i5}	P_{i6}	P_{i7}	P_{i8}	P_{i9}	P_{i10}	P_{i11}
						D = 0.45						
4	1	0.400	0.325	0.150	0.125	.	.	.	.	.	.	.
	2	0.125	0.150	0.325	0.400	.	.	.	.	.	.	.
7	1	0.050	0.100	0.125	0.250	0.225	0.150	0.100	.	.	.	.
	2	0.325	0.250	0.150	0.100	0.075	0.060	0.040	.	.	.	.
11	1	0.050	0.100	0.250	0.225	0.100	0.075	0.060	0.050	0.040	0.030	0.020
	2	0.045	0.050	0.055	0.060	0.065	0.100	0.200	0.175	0.125	0.075	0.050
						D = 0.65						
4	1	0.475	0.350	0.125	0.050	.	.	.	.	.	.	.
	2	0.050	0.125	0.350	0.475	.	.	.	.	.	.	.
7	1	0.025	0.050	0.100	0.250	0.250	0.200	0.125	.	.	.	.
	2	0.400	0.300	0.125	0.075	0.060	0.025	0.015	.	.	.	.
11	1	0.025	0.030	0.035	0.040	0.045	0.100	0.150	0.175	0.200	0.125	0.075
	2	0.275	0.225	0.175	0.100	0.050	0.040	0.035	0.030	0.025	0.025	0.020

TABLE 4.2

RESULTS OF MONTE CARLO SIMULATION STUDY: THE INDEX OF DISSIMILARITY IN A 2 X C TABLE WITH INDEPENDENT MULTINOMIAL ROW DISTRIBUTIONS

C	N_1	N_2	PREDICTED MEAN	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED DIFFERENCE	STANDARDIZED BIAS
				D = 0.05			
4	100	100	0.105283	0.106540	0.00204683	0.027793	1.249727
4	100	200	0.093470	0.093870	0.00156872	0.010104	1.107629
4	200	100	0.093470	0.093040	0.00162011	-0.010678	1.069302
4	200	200	0.080013	0.079490	0.00120994	-0.015024	0.847799
7	100	100	0.143567	0.141830	0.00184635	-0.040430	2.137113
7	100	200	0.126310	0.127010	0.00152951	0.017905	1.969115
7	200	100	0.126196	0.126465	0.00145558	0.007061	2.004217
7	200	200	0.106186	0.105105	0.00108756	-0.032768	1.670951
11	100	100	0.172118	0.169970	0.00212670	-0.046579	2.601475
11	100	200	0.150721	0.151370	0.00142197	0.017219	2.688213
11	200	100	0.150528	0.150110	0.00146454	-0.010931	2.615936
11	200	200	0.125537	0.125215	0.00092068	-0.010596	2.478851

TABLE 4.2 (CONTINUED)

C	N_1	N_2	PREDICTED MEAN	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED DIFFERENCE	STANDARDIZED BIAS
				D = 0.25			
4	100	100	0.257550	0.256760	0.00361770	-0.013142	0.112391
4	100	200	0.254450	0.255260	0.00278038	0.015355	0.099755
4	200	100	0.254450	0.256225	0.00310677	0.031839	0.111682
4	200	200	0.251805	0.251955	0.00213705	0.003239	0.042290
7	100	100	0.265936	0.266490	0.00382178	0.008956	0.266740
7	100	200	0.260155	0.259320	0.00292704	-0.015426	0.172267
7	200	100	0.259398	0.259110	0.00303536	-0.005235	0.165354
7	200	200	0.254462	0.253335	0.00198915	-0.025263	0.074776
11	100	100	0.284854	0.285440	0.00332961	0.010160	0.614182
11	100	200	0.273976	0.273550	0.00250485	-0.008504	0.470544
11	200	100	0.273201	0.274175	0.00269399	0.018756	0.465766
11	200	200	0.262884	0.264250	0.00166129	0.033517	0.349617

TABLE 4.2 (CONTINUED)

C	N_1	N_2	PREDICTED MEAN	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED DIFFERENCE	STANDARDIZED BIAS
				D = 0.45			
4	100	100	0.450050	0.451260	0.00432661	0.018398	0.019156
4	100	200	0.450009	0.446730	0.00289886	-0.060902	-0.060734
4	200	100	0.450009	0.448780	0.00322941	-0.021627	-0.021468
4	200	200	0.450000	0.448800	0.00203796	-0.026587	-0.026582
7	100	100	0.460484	0.458560	0.00371073	-0.031577	0.140522
7	100	200	0.457636	0.458205	0.00266780	0.011017	0.158855
7	200	100	0.457634	0.459560	0.00303066	0.034988	0.173656
7	200	200	0.454790	0.455255	0.00180051	0.010955	0.123844
11	100	100	0.474409	0.474290	0.00305950	-0.002147	0.439140
11	100	200	0.467661	0.467375	0.00247063	-0.005752	0.349559
11	200	100	0.468031	0.465810	0.00243109	-0.045052	0.320650
11	200	200	0.461303	0.461465	0.00182573	0.003785	0.268322

TABLE 4.2 (CONTINUED)

C	N_1	N_2	PREDICTED MEAN	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED DIFFERENCE	STANDARDIZED BIAS
				D = 0.65			
4	100	100	0.650001	0.650330	0.00253859	0.006521	0.006550
4	100	200	0.650000	0.650215	0.00225373	0.004526	0.004529
4	200	100	0.650000	0.651490	0.00210063	0.032507	0.032510
4	200	200	0.650000	0.649950	0.00141830	-0.001328	-0.001328
7	100	100	0.658048	0.658700	0.00261071	0.012763	0.170271
7	100	200	0.655847	0.655935	0.00185910	0.002048	0.137648
7	200	100	0.656245	0.657470	0.00221550	0.026026	0.158703
7	200	200	0.653850	0.652935	0.00129171	-0.025472	0.081663
11	100	100	0.661508	0.660830	0.00246621	-0.013652	0.218079
11	100	200	0.658955	0.660715	0.00193801	0.039978	0.243396
11	200	100	0.658896	0.658000	0.00185305	-0.020811	0.185843
11	200	200	0.656401	0.656255	0.00140900	-0.003887	0.166637

particularly when D is large and C is small. Specifically, as equations 4.24 and 4.25 indicate, the bias of $\hat{D}$ decreases as D increases and increases with C. The largest bias of $\hat{D}$ observed in the Monte Carlo study, absolutely and relatively, occurs when D = 0.05 and C = 11.

The predicted mean of $\hat{D}$ is also calculated for each set of Monte Carlo trials, using equation 4.24 with the appropriate values of C, N_1, N_2, and the multinomial probabilities in table 4.1. These predicted expectations are presented in table 4.2 as well. To aid our assessment of this approximation based on the normal and folded-normal distributions, the difference between the Monte Carlo and predicted means relative to the Monte Carlo standard error of $\hat{D}$ has been computed for every entry in table 4.2. We can observe that, with occasional exceptions, the normal approximation to the expected value of $\hat{D}$ accurately represents the means of $\hat{D}$ attained in the Monte Carlo study. Again with some irregularities, the accuracy of the predicted mean increases as N_1 and N_2 increase, and, as the signs of standardized differences attest, the approimation for the mean of $\hat{D}$ does not appear to systematically understate or overstate the means observed for the simulated $\hat{D}$. Interestingly, there also appears to be no clear relationship between the aggreement of the predicted and Monte Carlo means and C, the number of columns of the table, a somewhat encouraging result given the extremely small probabilities assigned to the rows in several of the trials and the usual warnings about the suitability of the normal approximation in such circumstances.

Unfortunately, the attempts to model the Monte Carlo distribution of $\hat{D}$ must be assessed as failures. Like the situation observed

in the normal distribution case when OVL is near unity (or D is near zero), the distribution of the simulated $\hat{D}$ in the 2 X C table seems to bunch when D is small. The tendency of the distribution of $\hat{D}$ to concentrate toward zero is not, however, as severe as that observed earlier in the normal case. In addition, while the Monte Carlo distribution of $\hat{D}$ appears to become symmetric for D distant from zero, or when N_1 and N_2 are sufficiently large, normality is uniformly rejected. In none of the 48 sets of simulation trials does a Kolmogorov test for normality, using the Stephans (1974) pseudocritical points, indicate that the normal distribution serves as an adequate probability model for the sampling distribution of $\hat{D}$ in the 2 X C table. Further attempts to fit the folded-normal and standard-beta distributions to these simulation data are also rejected by the Kolmogorov test. When C, N_1, and N_2 are all small, the distribution of $\hat{D}$ becomes quite discrete, so no continuous probability model may suffice in such circumstances. Whether a continuous distribution can represent the behavior of D when C, N_1, and N_2 are large can only be addressed when N_1 and N_2 are much larger that the values considered here.

The Multivariate Hypergeometric Model of the 2 X C Table

Let us now assume that the column totals, n_j (j=1,...,C), as well as the row totals, N_1 and N_2, of the 2 X C table are fixed. Let $N_1 + N_2 = N$. If the cell counts of the table x_{ij} (1=1,2; j=1,...,C) are regarded as a realization of the random assignment of the N individuals to the cells of the table subject only to the constraints imposed by the fixed row and column totals, the distribution of the x_{ij} can be written

in terms of the multivariate hypergeometric distribution of either $x_{11},\ldots,x_{1C}$ or $x_{21},\ldots,x_{2C}$ (Bishop et al., 1975, pp. 450-52). Here we shall work with the first row and its distribution, given by

$$P(x_{11},\ldots,x_{1C}) = \frac{\prod_{j=1}^{C}\binom{n_j}{x_{1j}}}{\binom{N}{N_1}}, \quad 0 \le x_{1j} \le J_j:$$

$$j=1,\ldots,C; \ \sum_{j=1}^{C} x_{1j} = N_1; \tag{4.32}$$

where $J_j = \min(n_j,N_1)$. Then (Steyn, 1955; Bishop et al., 1975) it is known that

$$E(x_{1j}) = \frac{n_j N_1}{N}, \quad j=1,\ldots,C; \tag{4.33}$$

$$\mathrm{Var}(x_{1j}) = \frac{n_j N_1}{N}\left(1 - \frac{n_j}{N}\right)\left(\frac{N - N_1}{N - 1}\right), \quad j=1,\ldots,C; \tag{4.34}$$

and

$$\mathrm{Cov}(x_{1j},x_{1j'}) = -\frac{n_j n_{j'} N_1}{N^2}\left(\frac{N - N_1}{N - 1}\right),$$

$$j=1,\ldots,C;\ j'=1,\ldots,C;\ j\neq j'. \qquad (4.35)$$

Let us define the random variable $\tilde{D}$ for this model of the 2 X C table in the following way:

$$\tilde{D} = \frac{1}{2}\sum_{j=1}^{C}|d_j|\ , \qquad (4.36)$$

where here

$$d_j = \frac{x_{1j}}{N_1} - \frac{n_j - x_{1j}}{N_2} = \frac{Nx_{1j} - n_j N_1}{N_1(N - N_1)}\ . \qquad (4.37)$$

Then

$$E(\tilde{D}) = \frac{1}{2}\sum_{j=1}^{C}\mu_j\ , \qquad (4.38)$$

and

$$\mathrm{Var}(\tilde{D}) = \frac{1}{4}\left[\sum_{j=1}^{C}\sigma_j^2 + \sum_{j\neq j'}^{C}\sum\sigma_{jj'}\right], \tag{4.39}$$

where μ_j, σ_j^2, and $\sigma_{jj'}$ are again defined as in (4.12), (4.13), and (4.14), but d_j (j=1,...,C) is defined as in (4.37).

Now it follows from the probability model in equation 4.32 that the marginal distribution of any x_{1j} (j=1,...,C) is hypergeometric, with probability function

$$P(x_{1j}) = \frac{\binom{n_j}{x_{1j}}\binom{N - n_j}{N_1 - x_{1j}}}{\binom{N}{N_1}}, \quad 0 \leq x_{1j} \leq J_j;\ j=1,\ldots,C; \tag{4.40}$$

where J_j is defined as above. The joint distribution of x_{1j} and $x_{1j'}$ ($j \neq j'$) is bivariate hypergeometric, with probability function

$$P(x_{1j},x_{1j'}) = \frac{\binom{n_j}{x_{1j}}\binom{n_{j'}}{x_{1j'}}\binom{N - n_j - n_{j'}}{N_1 - x_{1j} - x_{1j'}}}{\binom{N}{N_1}},$$

$$0 \le x_{1j} \le J_j;\ 0 \le x_{1j'} \le J_{j'};\ x_{1j} + x_{1j'} \le N_1. \quad (4.41)$$

By definition, μ_j, the expectation of $|d_j|$, can be computed as

$$\mu_j = \sum_{x_{1j}=0}^{J_j} \left| \frac{Nx_{1j} - n_j N_1}{N_1(N - N_1)} \right| \cdot P(x_{1j}) \ . \quad (4.42)$$

Noting that $E(d_j) = 0$ $(j=1,\ldots,C)$, we obtain from equation 4.19 the following expression for σ_j^2, the variance of $|d_j|$:

$$\sigma_j^2 = \frac{n_j(N - n_j)}{N_1(N - N_1)(N - 1)} - \mu_j^2 \ . \quad (4.43)$$

Finally, $\sigma_{jj'}$, the covariance of $|d_j|$ and $|d_{j'}|$ $(j \neq j')$ is given by

$$\sigma_{jj'} = \sum_{j=0}^{J_j} \sum_{j'=0}^{J_{j'}} \left[\left| \frac{(Nx_{1j} - n_j N_1)(Nx_{1j'} - n_{j'} N_1)}{N_1^2 (N - N_1)^2} \right| \cdot P(x_{1j}, x_{1j'}) \right] - \mu_j \mu_{j'} . \quad (4.44)$$

Substitution of these expressions for μ_j, σ_j^2, and $\sigma_{jj'}$ into equations 4.38 and 4.39 provides the mean and variance of $\tilde{D}$.

Normal Approximation to the Expectation and Variance of $\tilde{D}$

The extensive computations required to calculate μ_j, σ_j^2, and $\sigma_{jj'}$, particularly $\sigma_{jj'}$, suggest that some simpler method be used to find the mean and variance of $\tilde{D}$. Cortese et al. (1976) adopt a binomial approximation to the summation in (4.42) to calculate the expected value of $\tilde{D}$. Both jackknife (Taeuber and Taeuber, 1976) and bootstrap (Kestenbaum, 1980) methods have been advocated for computing the variance of $\tilde{D}$. The procedure introduced here parallels the method used earlier to derive the approximate mean of $\hat{D}$ in the multinomial case. The $x_{11},\ldots,x_{1C}$ are presumed to follow, at least approximately, a multivariate normal distribution with mean vector specified by equation 4.33 and variance-covariance matrix specified by equations 4.34 and 4.35. It immediately follows from this assumption that the d_j (j=1,...,C) are distributed as C-variate normal and the $|d_j|$ (j=1,...,C) are distributed as C-variate folded-normal. The relationship between the normal and folded-normal distributions then permits the derivation of relatively simple expressions for the approximate mean and variance of $\tilde{D}$.

We begin by noting the expectiation of each d_j (j=1,...,C) is zero and that τ_j^2, the variance of each d_j, is

$$\tau_j^2 = \frac{n_j(N - n_j)}{N_1(N - N_1)(N - 1)} . \tag{4.45}$$

Thus the properties of the folded-normal distribution require that μ_j, the expected value of $|d_j|$, be given by

$$\mu_j \doteq \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \cdot \tau = \left[\frac{2n_j(N - n_j)}{\pi N_1(N - N_1)(N - 1)}\right]^{\frac{1}{2}} . \tag{4.46}$$

The variance of $|d_j|$, σ_j^2, is simply

$$\sigma_j^2 \doteq \frac{\pi - 2}{\pi} \tau^2 = \frac{(\pi - 2)n_j(N - n_j)}{\pi N_1(N - N_1)(N - 1)} . \tag{4.47}$$

From equation 4.35, $\tau_{jj'}$, the covariance of d_j and $d_{j'}$ ($j \neq j'$), is

$$\tau_{jj'} = \frac{-n_j n_{j'}}{N_1(N - N_1)(N - 1)}, \tag{4.48}$$

and therefore the correlation between d_j and $d_{j'}$, $\rho_{jj'}$, is given by

$$\rho_{jj'} = -\left[\frac{n_j n_{j'}}{(N - n_j)(N - n_{j'})}\right]^{\frac{1}{2}}. \tag{4.49}$$

Hence by (4.30), the approximate covariance of $|d_j|$ and $|d_{j'}|$ $(j \neq j')$ is here

$$\sigma_{jj'} \doteq \frac{2[n_j(N - n_j)n_{j'}(N - n_{j'})]^{\frac{1}{2}}}{\pi N_1(N - N_1)(N - 1)}\left[(1 - \rho_{jj'}^2)^{\frac{1}{2}} + \right.$$

$$\left. + \rho_{jj'}\sin^{-1}(\rho_{jj'}) - 1\right]. \tag{4.50}$$

Combining these expressions for approximations to μ_j, σ_j^2, and $\sigma_{jj'}$, we obtain from (4.38) and (4.39) the following expressions for the approximate expectation and variance of $\tilde{D}$:

$$E(\tilde{D}) \doteq \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{C} \left[\frac{n_j(N - n_j)}{N_1(N - N_1)(N - 1)} \right]^{\frac{1}{2}}$$

$$= \left[2\pi(N - 1) \frac{N_1}{N} \left(1 - \frac{N_1}{N} \right) \right]^{-\frac{1}{2}} \cdot \sum_{j=1}^{C} \left[\frac{n_j}{N} \left(1 - \frac{n_j}{N} \right) \right]^{\frac{1}{2}}, \qquad (4.51)$$

and

$$\mathrm{Var}(\tilde{D}) \doteq \frac{1}{4} \left\{ \frac{\pi - 2}{\pi} \sum_{j=1}^{C} \frac{n_j(N - n_j)}{N_1(N - N_1)(N - 1)} + \right.$$

$$+ \frac{2}{\pi} \sum_{j \neq j'}^{C} \sum \frac{\left[n_j(N - n_j) n_{j'}(N - n_{j'}) \right]^{\frac{1}{2}}}{N_1(N - N_1)(N - 1)} \left[(1 - \rho_{jj'}^2)^{\frac{1}{2}} + \right.$$

$$\left. \left. + \rho_{jj'} \sin^{-1}(\rho_{jj'}) - 1 \right] \right\}$$

$$= \left[4\pi(N - 1) \frac{N_1}{N} \left(1 - \frac{N_1}{N} \right) \right]^{-1} \left\{ (\pi - 2) \left[1 - \sum_{j=1}^{C} \left(\frac{n_j}{N} \right)^2 \right] + \right.$$

$$+ 2 \sum_{j \neq j'}^{C} \sum \left[\frac{n_j}{N} \left(1 - \frac{n_j}{N} \right) \frac{n_{j'}}{N} \left(1 - \frac{n_{j'}}{N} \right) \right]^{\frac{1}{2}} \cdot \left[(1 - \rho_{jj'}^2)^{\frac{1}{2}} + \right.$$

$$\left. \left. + \rho_{jj'} \sin^{-1}(\rho_{jj'}) - 1 \right] \right\}. \qquad (4.52)$$

Clearly, both the expectation and variance of D appear to be functions of C, N, N_1/N, and n_j/N (j=1,...,C).

In the event that $n_1 = \ldots = n_C = n = N/C$, the expressions for

the approximate expectation and variance of $\tilde{D}$ simplify even further:

$$E(\tilde{D}) = (C - 1)^{\frac{1}{2}} \cdot \left[2\pi(N - 1) \frac{N_1}{N} \left(1 - \frac{N_1}{N} \right) \right]^{-\frac{1}{2}} , \tag{4.53}$$

and

$$\mathrm{Var}(\tilde{D}) \doteq \frac{C - 1}{C} \left(\pi + 2 \left\{ [C(C - 2)]^{\frac{1}{2}} - \sin^{-1} \left(\frac{-1}{C - 1} \right) - C \right\} \right) \times$$

$$\times \left[4\pi(N - 1) \frac{N_1}{N} \left(1 - \frac{N_1}{N} \right) \right]^{-1} . \tag{4.54}$$

Thus, more obviously than in the case of unequal column totals, we observe that $E(\tilde{D})$ increases with $C = N/n$, decreases as N increases, and is a quadratic function of N/N. While $\mathrm{Var}(\tilde{D})$ decreases as N increases, its dependence on C is more complex. We note, however, that as $C = N/n$ becomes large,

$$\lim_{C \to \infty} [\mathrm{Var}(\tilde{D})] = \left[4(N - 1) \frac{N_1}{N} \left(1 - \frac{N_1}{N} \right) \right]^{-1} .$$

Adequacy of the Normal Approximation to the Mean and Variance of $\tilde{D}$

Because the probability model for $\tilde{D}$, the index of dissimilarity in the 2 X C multivariate hypergeometric table, is so restrictive, we shall limit further consideration of the properties of $\tilde{D}$ to a comparison of the approximate expectation and variance of $\tilde{D}$ derived above to some Monte Carlo results in Kestenbaum (1980). Using the multivariate hypergeometric model, Kestenbaum generated simulated distributions for $\tilde{D}$ in the 2 X C table with equal column totals for several values of C, N, and N_1/N, and he computed the Monte Carlo mean and variance of the $\tilde{D}$ so obtained. The results Kestenbaum reports for N_1/N greater than or equal to 0.05 with N = 100 and N = 1000 are reproduced in table 4.3 and table 4.4. (Values of N_1/N less than 0.05 for these values of N seem patently unrealistic and are not considered here.) The values of C are 2, 4, and 10 for N = 100 and 10, 20, 40, and 100 for N = 1000. The Monte Carlo means and variances of $\tilde{D}$ in table 4.3 are computed from 1000 Monte Carlo trials at each combination of C and N_1/N; in table 4.4 the Monte Carlo means and variances are based on only 100 Monte Carlo trials. (Kestenbaum also reports Monte Carlo moments of $\tilde{D}$ when N = 10000, but these are computed from only 10 Monte Carlo trials.)

The approximate mean and variance of $\tilde{D}$ have been calculated from equations 4.53 and 4.54 using the appropriate values of C, N, and N_1/N; these are presented in tables 4.3 and 4.4 as the predicted mean and the predicted variance of $\tilde{D}$. Because Kestenbaum reports so few significant figures, the comparison of the predicted expectations and variances to his Monte Carlo results will necessarily be somewhat superficial. Nevertheless, the standardized difference--the Monte Carlo mean minus

TABLE 4.3

MEAN AND VARIANCE OF THE INDEX OF DISSIMILARITY IN THE MULTIVARIATE HYPERGEOMETRIC 2 X C TABLE WITH EQUAL COLUMN TOTALS (N = 100)

N_1/N	PREDICTED MEAN	PREDICTED VARIANCE	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED DIFFERENCE	VARIANCE RATIO
			C = 2			
0.05	0.183969	0.01931846	0.193	0.0158	0.0718	0.8179
0.10	0.133651	0.01019585	0.136	0.0107	0.0227	1.0494
0.20	0.100238	0.00573517	0.097	0.0062	-0.0411	1.0810
0.30	0.087495	0.00436965	0.088	0.0053	0.0069	1.2129
0.40	0.081844	0.00382345	0.081	0.0041	-0.0132	1.0723
0.50	0.080190	0.00367051	0.084	0.0041	0.0595	1.1170
			C = 4			
0.05	0.318644	0.01875999	0.322	0.0158	0.0267	0.8422
0.10	0.231490	0.00990110	0.242	0.0090	0.1108	0.9090
0.20	0.173617	0.00556937	0.175	0.0057	0.0183	1.0235
0.30	0.151546	0.00424333	0.154	0.0041	0.0383	0.9662
0.40	0.141758	0.00371291	0.141	0.0040	-0.0120	1.0773
0.50	0.138894	0.00356440	0.134	0.0027	-0.0942	0.7575

TABLE 4.3 (CONTINUED)

N_1/N	PREDICTED MEAN	PREDICTED VARIANCE	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED DIFFERENCE	VARIANCE RATIO
			C = 10			
0.05	0.551908	0.01908060	0.612	0.0059	0.7823	0.3092
0.10	0.400952	0.01007032	0.372	0.0115	-0.2700	1.1420
0.20	0.300714	0.00566455	0.285	0.0055	-0.2119	0.9710
0.30	0.262485	0.00431585	0.254	0.0048	-0.1225	1.1122
0.40	0.245532	0.00377637	0.237	0.0046	-0.1258	1.2181
0.50	0.240571	0.00362531	0.230	0.0038	-0.1715	1.0482

NOTE: MONTE CARLO MEANS AND VARIANCES ARE COMPUTED FROM 1000 MONTE CARLO TRIALS.

SOURCE: MONTE CARLO MEANS AND VARIANCES ARE FROM KESTENBAUM (1980).

TABLE 4.4

MEAN AND VARIANCE OF THE INDEX OF DISSIMILARITY IN THE MULTIVARIATE HYPERGEOMETRIC 2 X C TABLE WITH EQUAL COLUMN TOTALS (N = 1000)

N_1/N	PREDICTED MEAN	PREDICTED VARIANCE	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED DIFFERENCE	VARIANCE RATIO
			C = 10			
0.05	0.173741	0.00189087	0.175	0.0020	0.0282	1.0577
0.10	0.126220	0.00099796	0.128	0.0010	0.0563	1.0020
0.20	0.094665	0.00056135	0.091	0.0004	-0.1832	0.7126
0.30	0.082630	0.00042770	0.082	0.0004	-0.0315	0.9352
0.40	0.077293	0.00037423	0.076	0.0004	-0.0647	1.0688
0.50	0.075732	0.00035927	0.078	0.0004	0.1134	1.1134
			C = 20			
0.05	0.252440	0.00190259	0.260	0.0017	0.1834	0.8935
0.10	0.183393	0.00100414	0.180	0.0009	-0.1131	0.8963
0.20	0.137545	0.00056483	0.139	0.0005	0.0651	0.8852
0.30	0.120059	0.00043035	0.122	0.0003	0.1121	0.6971
0.40	0.112305	0.00037655	0.114	0.0004	0.0848	1.0623
0.50	0.110036	0.00036149	0.112	0.0004	0.0982	1.1065

TABLE 4.4 (CONTINUED)

N_1/N	PREDICTED MEAN	PREDICTED VARIANCE	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED DIFFERENCE	VARIANCE RATIO
			C = 40			
0.05	0.361670	0.00190851	0.364	0.0015	0.0602	0.7860
0.10	0.262747	0.00100727	0.268	0.0008	0.1857	0.7942
0.20	0.197061	0.00056659	0.196	0.0006	-0.0433	1.0590
0.30	0.172009	0.00043169	0.174	0.0004	0.0996	0.9266
0.40	0.160899	0.00037773	0.158	0.0004	-0.1450	1.0590
0.50	0.157648	0.00036262	0.158	0.0004	0.0176	1.1031

TABLE 4.4 (CONTINUED)

N_1/N	PREDICTED MEAN	PREDICTED VARIANCE	MONTE CARLO MEAN	MONTE CARLO VARIANCE	STANDARDIZED DIFFERENCE	VARIANCE RATIO
			C = 100			
0.05	0.576233	0.00191207	0.630	0.0004	2.6884	0.2092
0.10	0.418624	0.00100915	0.383	0.0014	-0.9521	1.3873
0.20	0.313968	0.00056765	0.300	0.0005	-0.6247	0.8808
0.30	0.274053	0.00043249	0.266	0.0004	-0.4027	0.9249
0.40	0.256354	0.00037843	0.252	0.0004	-0.2177	1.0570
0.50	0.251174	0.00036329	0.248	0.0004	-0.1587	1.1010

NOTE: MONTE CARLO MEANS AND VARIANCES ARE COMPUTED FROM 100 MONTE CARLO TRIALS.

SOURCE: MONTE CARLO MEANS AND VARIANCES ARE FROM KESTENBAUM (1980).

the predicted mean divided by the Monte Carlo standard error--indicates that the normal approximation to the mean of $\tilde{D}$ adequately represents the mean of the simulated $\tilde{D}$ when N_1/N is sufficiently large or C is sufficiently small. We can see that the predicted mean appears to lie closer to the Monte Carlo mean when N = 1000 (table 4.4) than when N = 100 (table 4.3). There is also some evidence in table 4.3 (C = 10) that when C is large relative to N, or n is small relative to N, the predicted mean overstates the expectation of $\tilde{D}$ observed in the Monte Carlo trials.

A result of perhaps more interest, since Cortese et al. (1976) have apparently developed an adequate approximation method for the expected value of $\tilde{D}$, is the comparison of the Monte Carlo variance given by Kestenbaum to that predicted by equation 4.54. As the ratio of these variances (Monte Carlo to predicted variance) demonstates, the predicted variance appears to overstate the Monte Carlo variance of $\tilde{D}$ for small N_1/N but is accurate, at least as far as Kestenbaum's results permit, when N_1/N is sufficiently large, where the N_1/N required increases as C increases. This, of course, is only to be expected, as the approximation formulae of equation 4.53 and equation 4.54 depend on the adequacy of the normal approximation to the multivariate hypergeometric distribution and thus, allowing for the usual requirements of such approximations, on C, since C is a function of the column totals of the 2 X C table and to the expected values of the $x_{11},\ldots,x_{1C}$.

Discussion

The behavior of $\hat{D}$ and $\tilde{D}$ as measures of association in the 2 X C contingency table indicate that the corresponding estimators of OVL, $\hat{OVL}$

and $\tilde{OVL}$, display the same properties as the estimators of the overlapping coefficient between two continuous distributions. The bias of $\hat{D}$ and $\tilde{D}$ demontrates that in the 2 X C table, estimators of OVL will exhibit downward bias and this bias is related to C, N_1, N_2, and the multinomial row probabilities in the multinomial case and to C, N, N_1, and the column totals n_j (j=1,...,C) in the multivariate hypergeometric case. The sampling variance of these estimators of OVL will be identical to the variance of $\hat{D}$ and $\tilde{D}$, so the relationships between the variance of the index of dissimilarity and the parameters of the assumed distribution of the 2 X C table obviously hold for the variance of the estimator of OVL under these probability models.

Of the two cases examined here, the multinomial model of the 2 X C table appears to be more relevant to our general exploration of the properties of OVL as a measure of agreement between distributions, for it corresponds to the comparison of two distributions through the arrangement of two independent samples from these distributions in the 2 X C table format. Once again, the evident bias of the estimator of the overlapping coefficient may be the most important property of $\hat{OVL}$ uncovered here. As in the case of the estimator of OVL between two normal distributions and the case of the spline-based estimator of OVL between two unspecified distributions, the closer the true overlap to unity, the greater the downward bias of $\hat{OVL}$. The fact that the mean, and in one circumstance the variance, of $\hat{D}$ and $\hat{OVL}$ can be closely approximated may prove useful in some applications. Several attempts to estimate the variance of $\hat{OVL}$ (and $\hat{D}$) in the multinomial case by the jackknife method (Efron, 1982, chap. 3) demonstate that the bootstrap

provides a better nonparametric estimate of the variance of $\hat{OVL}$. Apparently because of the discrete divisions of the 2 X C table, the jackknife substantially understated the Monte Carlo variance in every case examined. The bootstrap estimator of the variance of $\hat{OVL}$ in the 2 X C table is illustrated in the example below.

The primary importance of the multivariate hypergeometric model of the 2 X C table is the role is has assumed in the sociological and demographic literature. Since a part of the debate over the proper interpretation of the index of dissimilarity centers on its expectation and variance under this probability model, the approximate moments derived for $\tilde{D}$ may be helpful. First, the agreement between the approximate moments and the Monte Carlo moments reported by Kestenbaum suggests that in realistic applications, where N, N_1, and n_j (j=1,...,C) are large, the approximate moments will adequately represent the behavior of the random variable $\tilde{D}$ under the hypergeometric model, in so far as the mean and variance of $\tilde{D}$ describe this behavior. The fact that unequal column totals can be handled as easily as equal column totals in the approximation formulae indicates that these results have direct practical application. Second, the equations for the approximate mean and variance of $\tilde{D}$ clearly demonstrate the dependence of the behavior of $\tilde{D}$ on the column totals $n_1,\ldots,n_C$ as well as C, N, and N_1, an obvious feature which is sometimes overlooked when the multivariate hypergeometric model is introduced.

Whether the hypergeometric model is useful in the situations for which it is advocated is another question. It is obvious from the nonnormal distribution of $\hat{D}$ in the multinomial case that the

distribution of $\tilde{D}$ is certainly nonnormal as well, and thus the proposal of Cortese et al. (1976) that standardizing $\tilde{D}$ with respect to its mean and standard error does not seem compelling as an argument for overcoming the perceived limitations of $\tilde{D}$ itself. (In this regard, see also Cohen et al., 1976; Massey, 1978; Cortese et al., 1978; Falk et al., 1978.) Winship's objection to $\tilde{D}$ is that it is not sufficiently sensitive, since realizations of the 2 X C table which seem to indicate differing degrees of equity between the row distributions may yield the same value of $\tilde{D}$ (Winship, 1978). However, if we are willing to assume the multivariate hypergeometric model for the 2 X C table, an alternative to the use of $\tilde{D}$ (or any such measure of association) is the natural extension of Fisher's "exact" treatment of independence in the 2 X 2 table with fixed margins (Kendall and Stuart, 1979, pp. 580-83). That is, we simply compute the probability of obtaining the realization of the 2 X C table actually observed or one more extreme, using the probability function in equation 4.32 or some approximation to it (Freeman and Halton, 1951). The probability obtained has a natural interpretation, and it may prove more useful that $\tilde{D}$ wherever the multivariate hypergeometric model is reasonable.

An Example

As an example of the use of OVL in the 2 X C contingency table, we shall again use the sample of Alabama farm operators from the 1850 manuscript census described in Chapter Two. In the analysis of this sample subsequent to Inman (1981), a difference in the age distributions of slaveholders (N = 251) and nonslaveholders (N = 350) became apparent. Here the difference in these age distributions is examined by comparing

the ages of the slaveholders and the nonslaveholders in the sample, using the standard age categories favored by demographers; see table 4.5. While the majority of the nonslaveholders in the sample were in their twenties and thirties in 1850, the bulk of the slaveholders in the sample are spread in age from the late twenties through the early fifties. The value of the chisquare statistic computed from table 4.5 is 26.738, which with 11 degrees of freedom is statistically significant ($p = 0.0029$).

The overlap between the age distributions of these two classes of Alabama farmers estimated from this table is $\hat{OVL} = 0.832043$. To obtain a bootstrap estimate of the standard error of this estimated overlap, the RANTBL function in SAS (SAS, 1982) was used to generate 500 bootstrap $\hat{OVL}^*$ from table 4.5. Using these $\hat{OVL}^*$ and equations 3.19 and 3.20 we compute $\overline{\hat{OVL}^*} = 0.811419$ and $\hat{Var}_{500}(\hat{OVL}) = 0.00114277$; thus the bootstrap standard error for $\hat{OVL}$ in this example is 0.033805. From the bootstrap distribution function constructed from the 500 $\hat{OVL}^*$, a 90% confidence interval for OVL by the percentile method is (0.753773, 0.864200); see figure 4.1. The 90% bias-corrected confidence interval for the true overlap between the age distributions of Alabama slaveholding and nonslaveholding farmers in 1850 is (0.766090,0.871998); see figure 4.2. The computation of $\hat{OVL}$, the generation of the 500 $\hat{OVL}^*$, the calculation of the bootstrap variance of $\hat{OVL}$, and the construction of the bootstrap distribution function in SAS required less than one minute of CPU time on an IBM 4381-2.

Once more, the conclusion we should reach in this example is not that the age distributions are the same; we have already determined that

TABLE 4.5

AGE DISTRIBUTION OF ALABAMA FARMERS IN 1850

	AGE IN YEARS											
	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-79	TOTAL
NONSLAVEHOLDERS	40	58	57	47	39	36	35	19	6	5	8	350
SLAVEHOLDERS	7	28	34	40	39	31	33	16	10	7	6	251
TOTAL	47	86	91	87	78	67	68	35	16	12	14	601

SOURCE: INMAN (1981).

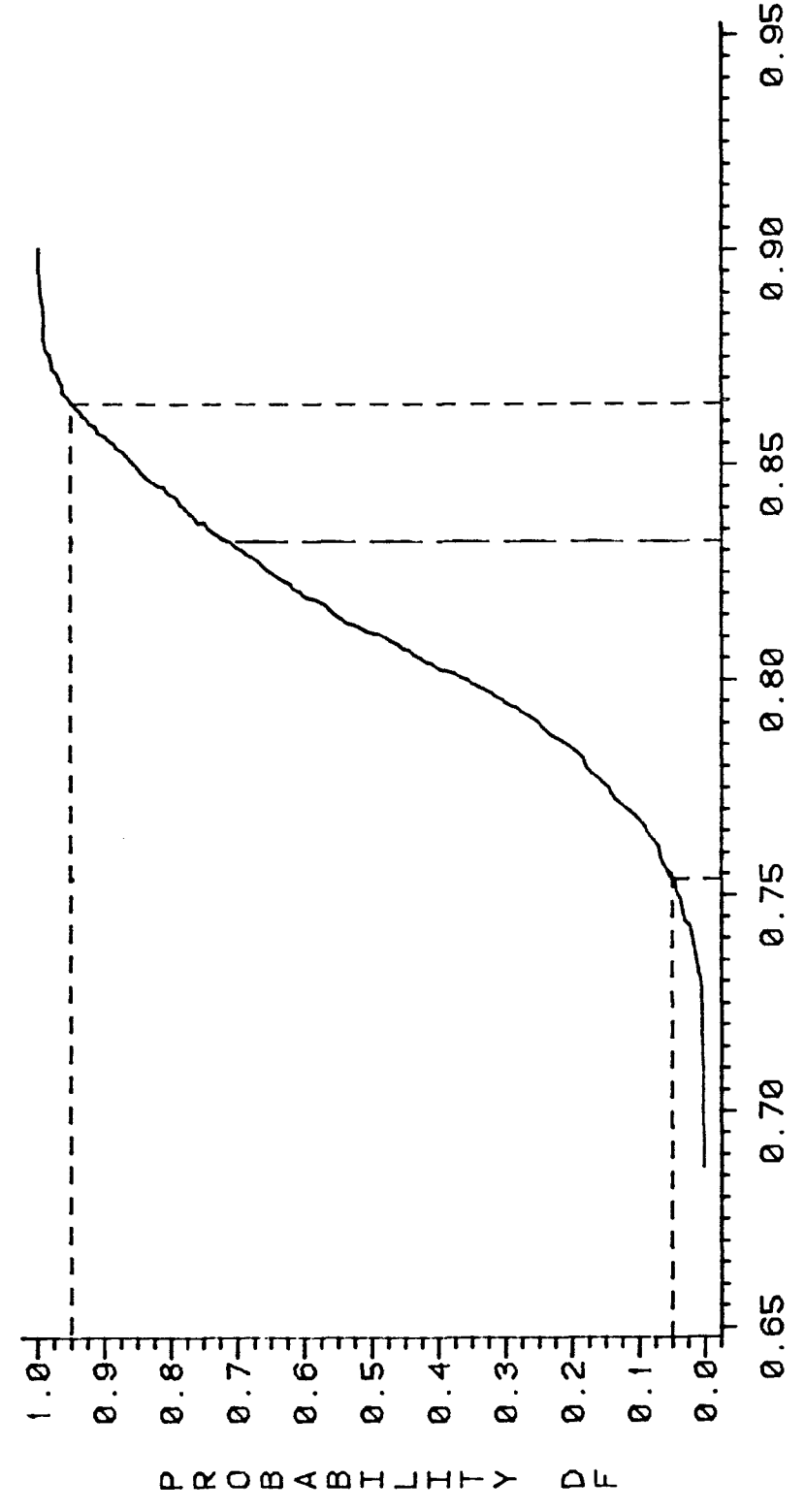

Figure 4.1 Construction of a 90% confidence interval for the overlap between the age distributions for slaveholding and nonslaveholding Alabama farmers in 1850 by the percentile method. The bootstrap distribution function for the estimator of OVL in table 4.5 (B = 500) is shown by the solid line. The estimated OVL is indicated by the heavy broken line, and the limits of the confidence interval are shown by the lighter broken lines.

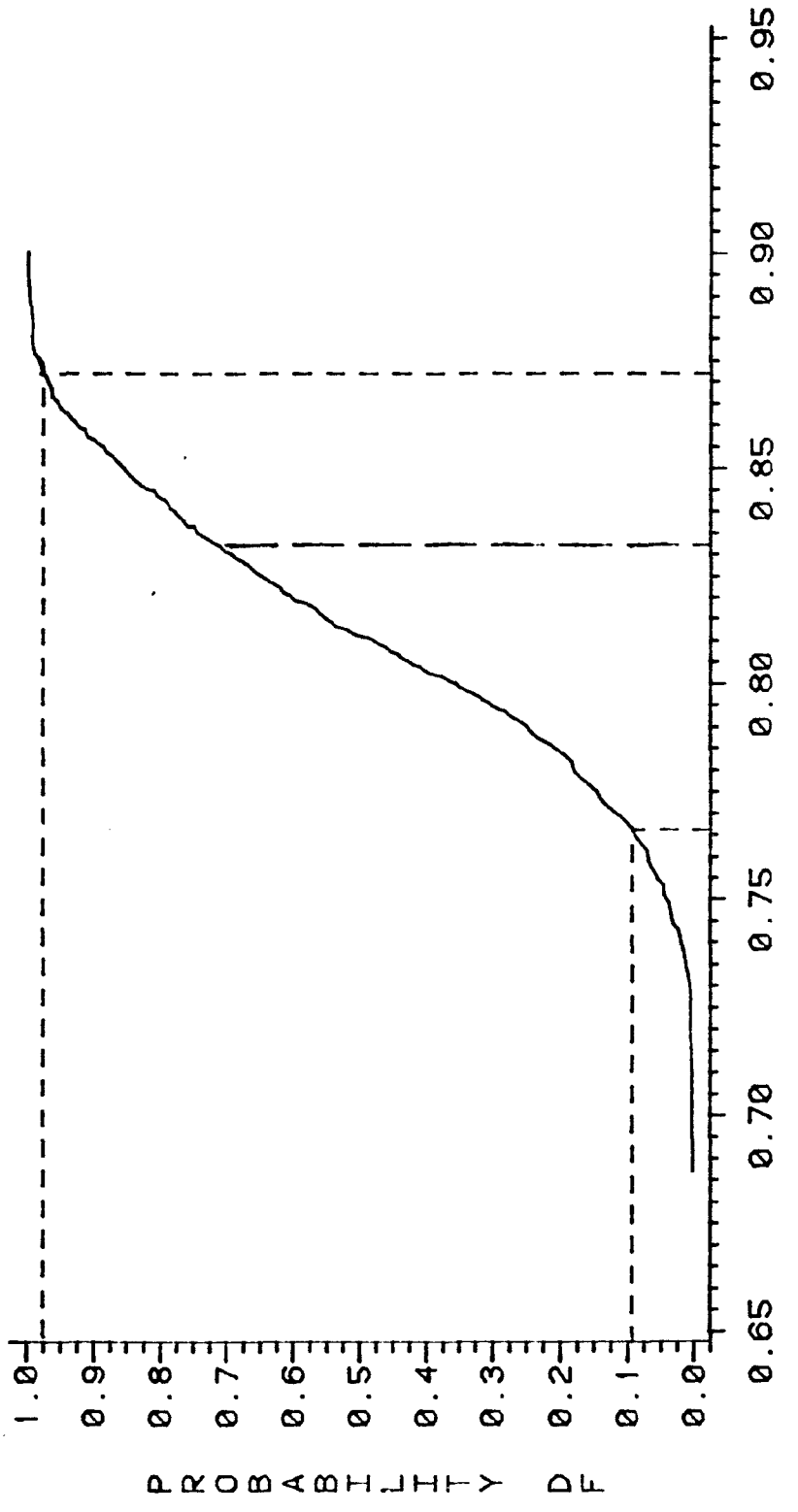

Figure 4.2 Construction of a 90% confidence interval for the overlap between the age distributions for slaveholding and nonslaveholding Alabama farmers in 1850 by the bias-corrected percentile method. The bootstrap distribution function for the estimator of OVL in table 4.5 (B = 500) is shown by the solid line. The estimated OVL is indicated by the heavy broken line, and the limits of the confidence interval are indicated by the lighter broken lines.

the distributions differ. Instead, OVL provides an indication of the importance of the difference between the two age distributions. In this example, $\hat{OVL}$ actually represents the common area under the two age distributions of interest, where these distributions are estimated by the relative frequency histograms summarized in table 4.5. Inspecting these histograms, shown in figure 4.3, visually reinforces the message of $\hat{OVL}$: The age distributions of the farmers who did and did not own slave property in 1850 are different; this we should expect, given the general association of wealth and age in the nineteenth-century United States. Too narrow a focus on the difference in the ages of slaveholding and nonslaveholding Alabama farmers, however, misses the considerable similarity of the two age distributions.

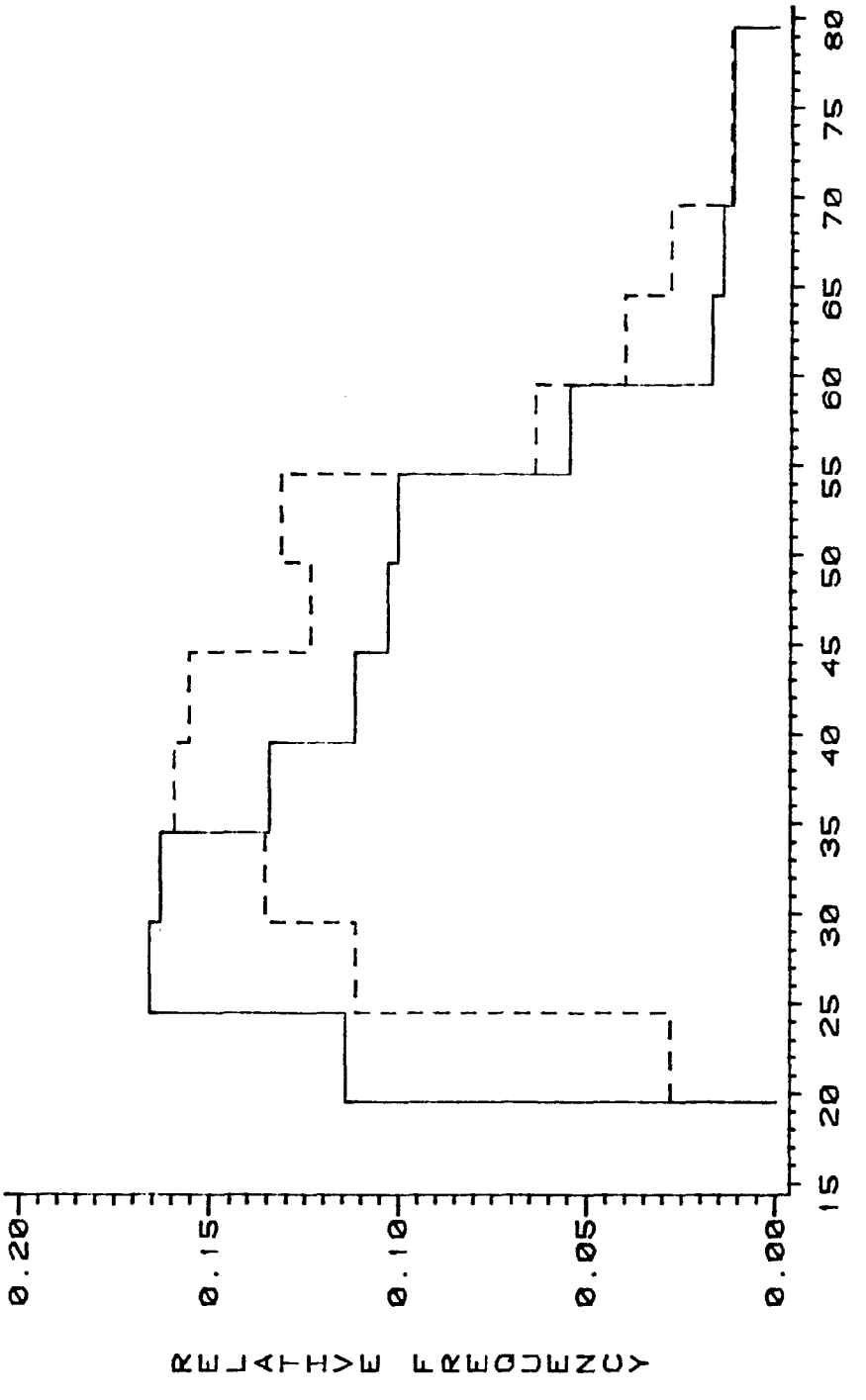

Figure 4.3 Relative frequency histograms for the age of Alabama farm operators in 1850. The age distribution for nonslaveholders is shown by the solid line; the age distribution for slaveholders is indicated by the broken line.

Chapter Five

CONCLUSION

The implications of this study of the behavior of the overlapping coefficient can be summarized briefly. Although the sample estimators of OVL investigated in the three distributional settings investigated here are consistent estimators of the true overlap between the distributions from which the samples are obtained, the downward bias of these estimators and the relationship of this bias to the true overlap suggest that sample estimators of OVL are not useful as test statistics for the equality of the two distributions being compared. Since in each of the distributional settings examined here there are accepted statistical techniques for testing the equality of the distributions of interest, this result should not necessarily disturb us. Likelihood ratio tests of the usual tests for the equality of means and variances of two normal populations, Kolomogorov-like tests for the equality of two unspecified distributions using the sample distribution functions, and various tests related to the chisquare statistic speak directly to the problem of whether two unknown distributions are identical in the normal distribution case, the nonparametric case, and the 2 X C contingency table case respectively. What OVL can provide is some measure of the meaningfulness of the differences that are detected by these statistical techniques. Thus OVL offers one method of exploring the

practical significance of differences which appear to be statistically significant.

As a sample statistic, OVL suffers not only from its bias but from the difficulties associated with estimating its standard error as well. The situation where estimates of the variance of the sample overlapping coefficient appear reliable, the 2 X C multivariate hypergeometric table case, is so restricted that it is unlikely to be of major importance. In the normal distribution case, the estimators of the sampling variance derived by statistical differentials are likely to prove reliable in limited circumstances also. Therefore the bootstrap method of obtaining the estimated standard error of the estimators of OVL will undoubtedly serve in practice as the most effective procedure for both ascertaining the standard error of the estimator of OVL and constructing confidence intervals for OVL between the distributions of interest.

Why, then, use OVL at all? In the case of two normal distributions with equal variances, a components of variance approach addresses the same issue, and even the simple characterization of the magnitude of the difference between population means in units of the population standard deviation may suffice (Cohen, 1977). In the case of the 2 X C table, measures of association exist in bountiful supply, and OVL may be no more attractive than many of others.

The advantage of OVL is two-fold. First, it offers a common approach for the measurement of the agreement between two distributions in any distributional setting. In this sense, then, OVL is less restrictive than other procedures keyed directly to distributional

assumptions that may or may not prove warranted in data analysis. Second, OVL is based on a simple, easily comprehended concept of the association between distributions. While the simplicity of OVL as the ammount of probability mass common to both distributions is appealing in its own right, OVL also has another interpretation based on the classification of individuals into two populations. Given the two distributions representing the populations of interest, OVL represents the sum of the conditional probabilities of misclassifying an individual into the two populations, where the classification rule is the assignment of an individual at any level of the characteristic of concern to the population eith the greater probability at that level. Thus OVL can be regarded as an indicator of the difference between individuals in the two populations or distributions of interest. Whether or not OVL proves useful in any given setting, then, depends on the meaning OVL has in the context of the specific problem and the value of its general approach. The fact that the problem it addresses, the meaningfulness of differences between the two distributions of interest, is raised may be more important that whether the overlapping coefficient is adopted as a possible solution.

APPENDIX

FORTRAN SUBROUTINES

The following FORTRAN subroutines were used in the Monte Carlo simulation studies of the overlapping coefficient described in Chapter Two and Chapter Three. They are presented here as documentation, not beacause the code is particularly innovative. Some of the subroutines are called from others; in several of the subroutines, FORTRAN routines available from de Boor (1978), Hanson (1979), and IMSL (1982) are called. Calls to such subroutines are noted in the prefaces to each of the following subroutines.

Subroutine BSPLDF

The object of subroutine BSPLDF is to obtain by weighted least squares the quadratic spline estimate of an unknown distribution function from a sample distribution function. The Hanson (1979) routine FC is used to obtain the coefficients of the quadratic B-spline.

Called subroutines: BSPLPP (de Boor), FC (Hanson), NEWNOT (de Boor), RSSQDF, and XSETF (Hanson).

```
      SUBROUTINE BSPLDF(NDATA,XDATA,YDATA,SDDATA,NORD,BKLOW,BKUP,IPASS,N
     1BKPT,BKPT,NCOEFF,COEFF,RSSQ,MODE)
C
C     SUBROUTINE BSPLDF CALCULATES THE BREAKPOINTS AND COEFFICIENTS OF
C     A B-SPLINE ESTIMATED (CUMULATIVE) DISTRIBUTION FUNCTION FOR THE
C     EMPIRICAL CDF CONTAINED IN THE ARRAYS XDATA, YDATA.  NDATA IS THE
C     NUMBER OF DISTINCT POINTS IN THE EMPIRICAL CDF.  XDATA CONTAINS
C     THE NDATA POINTS OF THE VARIABLE X.  YDATA CONTAINS THE ESTIMATED
C     VALUE OF THE DISTRIBUTION FUNCTION AT THOSE POINTS.  SDDATA CON-
C     TAINS THE ESTIMATES OF THE STANDARD ERROR OF THE CDF AT EACH POINT
C     IN XDATA.  NORD IS THE ORDER OF THE B-SPLINE (NORD=ORDER+1).
C     NINT IS THE NUMBER OF INTERVALS -- DECILES (NINT=10), FOR EXAMPLE
C     -- USED TO OBTAIN THE BREAKPOINTS OF THE B-SPLINE.  BKUP IS A
C     POINT TO THE RIGHT OF ALL POINTS IN XDATA BY WHICH THE
C     DISTRIBUTION FUNCTION IS ASSUMED TO EQUAL ONE; BLOW IS A SIMILAR
C     POINT TO THE LEFT OF ALL POINTS IN XDATA AT WHICH THE DISTRIBUTION
C     FUNCTION IS ASSUMED TO BE EQUAL TO ZERO.  THE ROUTINE BSPLDF()
C     RETURNS THE NUMBER OF BREAKPOINTS (INCLUDING THOSE CREATED TO THE
C     LEFT AND RIGHT OF THE POINTS BKLOW AND BKUP TO FIT THE B-SPLINE),
C     THE ARRAY BKPT CONTAINING THE BREAKPOINT VALUES, AND COEFF, THE
C     ARRAY OF B-SPLINE COEFFICIENTS.  NOTE THAT NBKPT=NINT+2*NORD-1,
C     AND THAT THE NUMBER OF COEFFICIENTS IS NCOEFF=NBKPT-NORD.
C     MODE IS THE HANSON DIAGNOSTIC VARIABLE.
C
C     NOTE THE NUMBER OF INTERVALS USED FOR CONSTRUCTION OF THE B-SPLINE
C     DISTRIBUTION FUNCTION (AND DENSITY), NINT, IS BASED ON STURGES'S
C     RULE FOR THE NUMBER OF INTERVALS IN A HISTOGRAM.  SEE H. A.
C     STURGES, JASA 21 (1926): 65-66.
C
C     AFTER THE ROUTINE FC() OBTAINS  INITIAL ESTIMATES OF THE B-SPLINE
C     USING THE BREAKPOINTS GENERATED FROM THE EMPIRICAL CDF QUANTILES,
C     THE DE BOOR ROUTINE NEWNOT() IS USED TO OBTAIN A NEW SET OF BREAK-
C     POINTS AND THE B-SPLINE IS REESTIMATED .  THE VARIABLE IPASS SETS
C     THE NUMBER OF PASSES THROUGH THE PROCEDURE NEWNOT().
C
      DIMENSION XDATA(NDATA),YDATA(NDATA),SDDATA(NDATA)
      DIMENSION BKPT(40),XCONST(50),YCONST(50),NDERIV(50)
      DIMENSION W(10000),IW(250)
      DIMENSION COEFF(50)
      DIMENSION SCRTCH(20,20),PPCOEF(20,50),COEFG(20,50),PPBKPT(40)
      DIMENSION BKNEW(40)
C
      IKNOT=0
      NINT=IFIX(1.0+3.3*ALOG10(FLOAT(NDATA)))
      NINTP1=NINT+1
      NINTL1=NINT-1
      NBKPT=NINT+2*NORD-1
      NCOEFF=NBKPT-NORD
C
C     GET THE INITIAL BREAKPOINTS FOR THE B-SPLINE ROUTNIE FC().
C
      BKPT(NORD)=BKLOW
      BKPT(NORD+NINT)=BKUP
      K=0
      FNINT=FLOAT(NINT)
      DO 360 J=1,NINTL1
      F=FLOAT(J)/FNINT
  310 K=K+1
      D=YDATA(K)-F
```

```
      IF (D) 320,340,330
 320  DLAST=D
      GO TO 310
 330  DCOMP=ABS(D)-ABS(DLAST)
      IF (DCOMP) 340,350,350
 340  BKPT(NORD+J)=XDATA(K)
      GO TO 360
 350  BKPT(NORD+J)=XDATA(K-1)
 360  CONTINUE
C
C     EXTERIOR BREAKPOINTS SET EQUAL TO INTERVAL (BKLOW,BKUP) ENDPOINTS.
C
      NORDL1=NORD-1
      DO 380 J=1,NORDL1
       BKPT(J)=BKLOW
      BKPT(NORD+NINT+J)=BKUP
 380  CONTINUE
C
C     WRITE CONSTRAINTS FOR B-SPLINE ROUTINE.
C
      NCONST=NINT+3
C
C     CONSTRAIN B-SPLINE TO BE ZERO AT LEFT-MOST BREAKPOINT.
C
      XCONST(1)=BKPT(NORD)
      YCONST(1)=0.0
      NDERIV(1)=2
C
C     CONSTRAIN FIRST DERIVATIVE TO BE ZERO AT LEFT-MOST BREAKPOINT.
C
      XCONST(2)=BKPT(NORD)
      YCONST(2)=0.0
      NDERIV(2)=6
C
C     CONSTRAIN FIRST DERIVATIVE TO BE NONNEGATIVE AT ALL INTERIOR
C     BREAKPOINTS.
C
      DO 400 I=1,NINTL1
      J=I+2
      XCONST(J)=BKPT(NORD+I)
      YCONST(J)=0.0
      NDERIV(J)=5
 400  CONTINUE
C
C     CONSTRAIN B-SPLINE TO BE ONE AT RIGHT-MOST BREAKPOINT.
C
      XCONST(NINT+2)=BKPT(NORD+NINT)
      YCONST(NINT+2)=1.0
      NDERIV(NINT+2)=2
C
C     CONSTRAIN THE FIRST DERIVATIVE TO BE ZERO AT RIGHT-MOST
C     BREAKPOINT.
C
      XCONST(NINT+3)=BKPT(NORD+NINT)
      YCONST(NINT+3)=0.0
      NDERIV(NINT+3)=6
C
C     CALL HANSON ROUTINE FC() TO GET B-SPLINE COEFFICIENTS.
C
```

```
      IW(1)=10000
      IW(2)=250
      MODE=2
      CALL XSETF(2)
      CALL FC(NDATA,XDATA,YDATA,SDDATA,NORD,NBKPT,BKPT,NCONST,XCONST,YCO
     1NST,NDERIV,MODE,COEFF,W,IW)
C
C     BEGIN ITERATED CONSTRUCTION OF NEW KNOT SEQUENCES.
C
      IF (IKNOT .GE. IPASS) GO TO 440
      IKNOT=IKNOT+1
C
C     GET PP-REPRESENTATION OF B-SPLINE FOR NEWNOT PROCEDURE.
C
C     SCRTCH IS WORK SPACE DIMENSIONED (NORD,NORD).  NOTE OUTPUT OF
C     BSPLPP().  ARRAY PPBKPT CONTAINS THE PP-REP. BREAKPOINTS.  ARRAY
C     PPCOEF CONTAINS THE PP-REP. COEFFICIENTS.  L IS NINT, THE NUMBER
C     OF SUBINTERVALS INTO WHICH THE INTERVAL (BKLOW,BKUP) IS DIVIDED.
C     FOR DOCUMENTATION OF BSPLPP() SEE DE BOOR, PP. 140-41.
C
      CALL BSPLPP(BKPT,COEFF,NCOEFF,NORD,SCRTCH,PPBKPT,PPCOEF,L)
C
C     GET NEW SEQUENCE OF INTERIOR KNOTS.
C
C     NOTE OUTPUT OF NEWNOT().  ARRAY BKNEW CONTAINS (NINT+1) NEW
C     BREAKPOINTS, INCLUDING BKLOW AND BKUP.  BKNEW(1)=BKPT(NORD)=BKLOW,
C     AND BKNEW(NINT+1)=BKPT(NORD+NINT)=BKUP.  ARRAY COEFG CON-
C     TAINS THE COEFFICIENT PART OF THE PP-REPR. BKPT, COEFG, L, 2 FOR
C     THE MONOTONE P.LINEAR FUNCTION G  WRTO WHICH BKNEW WILL BE
C     EQUIDISTRIBUTED.
C     FOR DOCUMENTATION OF NEWNOT() SEE DE BOOR, PP.184-86.
C
      CALL NEWNOT(PPBKPT,PPCOEF,L,NORD,BKNEW,NINT,COEFG)
C
C     CONSTRUCT NEW SEQUENCE OF BREAKPOINTS FOR B-SPLINE ROUTINE FC().
C
      DO 420 I=1,NINTP1
      J=NORD+I-1
      BKPT(J)=BKNEW(I)
 420  CONTINUE
      GO TO 360
 440  CONTINUE
C
C     CALCULATE RESIDUAL SUM OF SQUARES FOR B-SPLINE DF.
C
      CALL RSSQDF(NDATA,XDATA,YDATA,NORD,BKPT,NBKPT,COEFF,NCOEFF,RSSQ)
      RETURN
      END
```

Real Function DNORML

The purpose of the real function DNORML is to evaluate the standard normal density function at the point x.

Called subroutines: none.

```
      REAL FUNCTION DNORML(X)
C
C     EVALUATE THE STANDARD NORMAL DENSITY AT X.
C
      IF (ABS(X) .GT. 13.0) GO TO 10
      DNORML=EXP(-X**2/2.0)/SQRT(6.2831853)
      GO TO 20
 10   DNORML=0.0
 20   CONTINUE
      RETURN
      END
```

Subroutine EMPCDF

The purpose of the subroutine EMPCDF is to construct the empirical distribution function from sample data.

Called subroutines: VSRTA (IMSL).

```
      SUBROUTINE EMPCDF(XWORK,NWORK,NCDF,XCDF,YCDF,SDCDF)
C
C     SUBROUTINE EMPCDF CALCULATES THE POINTS OF AN EMIRICAL CUMULATIVE
C     DISTRIBUTION FUNCTION FOR THE DATA ARRAY XWORK.  NWORK IS NUMBER
C     OF DATA POINTS IN THE ARRAY XWORK.  THE SUBROUTINE RETURNS NCDF,
C     THE NUMBER OF DISTINCT POINTS OF THE EMPIRICAL CDF; XCDF, THE
C     SORTED XDATA POINTS OF THE EMPIRICAL CDF; YCDF, THE ESTIMATED CDF
C     AT THE POINTS XCDF (USING THE DIVISOR (NWORK+1)); AND SDCDF, THE
C     ESTIMATED STANDARD ERROR OF THE EMPIRICAL CDF AT EACH POINT
C     IN XCDF.  THE MAXIMUM LENGTH OF THE ARRAYS XCDF, YCDF, AND SDCDF
C     IS NWORK.  EMPCDF() USES THE IMSL ROUTINE VSRTA TO SORT THE DATA
C     IN XWORK.
C
      DIMENSION XWORK(NWORK),XCDF(NWORK),YCDF(NWORK),SDCDF(NWORK)
      CALL VSRTA(XWORK,NWORK)
      NDIV=NWORK+1
      FNWORK=FLOAT(NWORK)
      FNDIV=FLOAT(NDIV)
      NDUP=0
      DO 180 I=2,NWORK
      J=I-1
      K=J-NDUP
      IF (XWORK(I) .EQ. XWORK(J)) GO TO 100
      XCDF(K)=XWORK(J)
      YCDF(K)=FLOAT(J)/FNDIV
      SDCDF(K)=SQRT(YCDF(K)*(1.0-YCDF(K))*FNWORK/FNDIV**2)
      GO TO 120
  100 NDUP=NDUP+1
  120 IF (I-NWORK) 180,160,140
  140 STOP
  160 NCDF=NWORK-NDUP
      XCDF(NCDF)=XWORK(NWORK)
      YCDF(NCDF)=FNWORK/FNDIV
      SDCDF(NCDF)=SQRT(YCDF(NCDF)*(1.0-YCDF(NCDF))*FNWORK/FNDIV**2)
  180 CONTINUE
      RETURN
      END
```

Subroutine NSTAT

The soubroutine NSTAT computes the maximum-likelihood estimates using the West algorithm (Chan and Lewis, 1979) of the mean and variance of a normal population from a simple random sample.

Called subroutines: none.

```
      SUBROUTINE NSTAT(X,N,U,V)
C
C     CALCULATE MAXIMUM LIKELIHOOD ESTIMATES OF THE NORMAL MEAN AND
C     VARIANCE FOR SAMPLE ARRAY X OF SIZE N USING WEST'S ALGORITHM.  U
C     IS THE SAMPLE MEAN, V IS THE SAMPLE VARIANCE.  SEE TONY F. CHAN
C     AND JOHN GREGG LEWIS, COMMUNICATIONS OF THE ACM 22 (SEPT. 1979):
C     528.
C
      DIMENSION X(N)
      SUMM=X(1)
      SUMT=0.0
      DO 10 I=2,N
      XDIF=X(I)-SUMM
      XDIFIX=XDIF/FLOAT(I)
      SUMM=SUMM+XDIFIX
      SUMT=SUMT+FLOAT(I-1)*XDIF*XDIFIX
 10   CONTINUE
      U=SUMM
      V=SUMT/FLOAT(N)
      RETURN
      END
```

Subroutine OVLEQ

The subroutine OVLEQ computes $\hat{OVL}$ and the estimated (approximate) variance of $\hat{OVL}$ for two normal distributions with equal population variances.

Called subroutines: DNORML, MDNOR (IMSL).

```
      SUBROUTINE OVLEQ(NONE,U1,V1,NTWO,U2,V2,IWRITE,OVL,VOVL)
C
C     CALCULATE THE OVERLAPPING COEFFICIENT AND THE VARIANCE OF ITS
C     SAMPLE ESTIMATOR FOR THE CASE OF SAMPLING FROM TWO NORMAL POPULA-
C     TIONS WITH EQUAL VARIANCES.  THE SUBROUTINE ASSUMES SAMPLE
C     ESTIMATES OF THE MEANS AND VARIANCES ARE INPUTTED FOR THE VALUES
C     OF U1 AND U2, THE MEANS OF THE TWO POPULATIONS, AND V1 AND V2, THE
C     SAMPLE ESTIMATES OF THE COMMON VARIANCE, AND CALCULATES A POOLED
C     ESTIMATE OF THE COMMON VARIANCE FROM V1 AND V2.  IF THIS VARIANCE
C     IS KNOWN, THEN THIS VALUE SHOULD BE USED FOR BOTH V1 AND V2 IN THE
C     CALL TO THE ROUTINE.
C
C     NOTE: IMSL ROUTINE MDNOR() IS USED TO EVALUATE THE STANDARD NORMAL
C     DISTRIBUTION FUNCTION.
C
C     CALCULATE THE POOLED ESTIMATE OF THE COMMON POPULATION VARIANCE.
C
      FNONE=FLOAT(NONE)
      FNTWO=FLOAT(NTWO)
      FNSUM=FLOAT(NONE+NTWO)
      VPOOL=(FNONE*V1+FNTWO*V2)/FNSUM
C
C     CALCULATE OVL.
C
      SIGMA=SQRT(VPOOL)
      UDIFF=U1-U2
      DELL=UDIFF/SIGMA
      Y1=-ABS(DELL/2.0)
      CALL MDNOR(Y1,P1)
      OVL=2.0*P1
C
C     CALCULATE THE VARIANCE OF THE SAMPLE ESTIMATOR OF OVL.
C
C     FIND THE EXPECTATION OF ABS(XBAR1-XBAR2).
C
      FACTOR=SQRT(FNSUM/(FNONE*FNTWO))
      DELFAC=DELL/FACTOR
      Y2=-DELFAC
      CALL MDNOR(Y2,P2)
      EXPECT=SQRT(0.6366198)*SIGMA*FACTOR*EXP(-DELFAC**2/2.0)+UDIFF*(1.0
     1   -2.0*P2)
C
C     COMPUTE VARIANCE OF OVLHAT.
C
      VOVL=(DNORML(Y1))**2*(FACTOR**2+DELL**2*(1.0+0.5*(FNSUM-2.0)/FNSUM
     1**2)-EXPECT**2/VPOOL)
C
      IF (IWRITE .EQ. 0) GO TO 50
      WRITE(6,40) OVL
 40   FORMAT(1H0,'THE OVERLAPPING COEFFICIENT:',F20.8)
      WRITE(6,41) EXPECT
 41   FORMAT(1H0,'EXPECTED VALUE OF THE ABSOLUTE DIFFERENCE IN SAMPLE ME
     1ANS:',F20.8)
      WRITE(6,44) VOVL
 44   FORMAT(1H0,'THE APPROXIMATE VARIANCE OF THE SAMPLE OVERLAPPING COE
     1FFICIENT:',F20.12)
 50   CONTINUE
      RETURN
      END
```

Subroutine OVLNEQ

The subroutine OVLNEQ computes $\hat{OVL}$ and its estimated (approximate) variance for two normal distributions with unequal population variances.

Called subroutines: DNORML, MDNOR (IMSL).

```
      SUBROUTINE OVLNEQ(N10,U10,V10,N20,U20,V20,IWRITE,OVL,VOVL)
C
C     CALCULATE THE (APPROXIMATE) VARIANCE OF OVLHAT, THE MAXIMUM
C     LIKELIHOOD ESTIMATE OF THE TRUE OVERLAP BETWEEN TWO NORMAL
C     DISTRIBUTIONS WITH UNEQUAL VARIANCES, USING AN APPROXIMATION
C     PROCEDURE BASED ON THE TECHNIQUE OF STATISTICAL DIFFERENTIALS.
C
C     IF IWRITE IS SET EQUAL TO ZERO, NO OUTPUT WILL BE PRINTED, UNLESS
C     V1 AND V2 ARE EQUAL.  IN THIS CASE A WARNING MESSAGE IS PRINTED
C     AND THE OUTPUTTED VALUES FOR OVL AND VOVL ARE BOTH SET TO ZERO.
C
C     NOTE:  V2 IS ASSUMED TO BE THE LARGER OF THE TWO VARIANCES, V1 AND
C     V2.  U1 AND U2 ARE THE MEANS ASSOCIATED WITH V1 AND V2
C     RESPECTIVELY.  IF V1 IS LARGER THAN V2, THE SAMPLE SIZES, MEANS,
C     AND VARIANCES ARE INTERCHANGED SO THAT V2 IS THE LARGER VARIANCE.
C
C     NOTE:  IMSL ROUTINE MDNOR() IS USED TO EVALUATE THE STANDARD
C     NORMAL DISTRIBUTION FUNCTION.
C
C     COMPARE VARIANCES.
C
      IF (V20-V10) 10,997,20
 10   NONE=N20
      NTWO=N10
      U1=U20
      U2=U10
      V1=V20
      V2=V10
      GO TO 30
 20   NONE=N10
      NTWO=N20
      U1=U10
      U2=U20
      V1=V10
      V2=V20
 30   CONTINUE
C
C     CALCULATE CROSSING POINTS.  X1 IS LOWER CROSSING POINT, X2 IS THE
C     UPPER CROSSING POINT.
C
      SD1=SQRT(V1)
      SD2=SQRT(V2)
      SD1SD2=SD1*SD2
      UDIFF=U1-U2
      VDIFF=V2-V1
      TERM1=U1*V2-U2*V1
      TERM2=SQRT(UDIFF**2+VDIFF*ALOG(V2/V1))
      X1=(TERM1-SD1SD2*TERM2)/VDIFF
      X2=(TERM1+SD1SD2*TERM2)/VDIFF
C
C     COMPUTE THE STANDARDIZED VALUES OF THE CROSSING POINTS.  ZIJ IS
C     XI STANDARDIZED WITH RESPECT TO MEAN UJ AND STANDARD DEVIATION
C     SDJ.
C
      Z11=(X1-U1)/SD1
      Z12=(X1-U2)/SD2
      Z21=(X2-U1)/SD1
      Z22=(X2-U2)/SD2
C
```

```
C     CALCULATE THE OVERLAPPING COEFFICIENT.
C
      CALL MDNOR(Z11,P11)
      CALL MDNOR(Z22,P22)
      CALL MDNOR(Z12,P12)
      CALL MDNOR(Z21,P21)
      OVL=1.0+P11+P22-P12-P21
C
C     COMPUTE VARIANCES OF THE MAXIMUM LIKELIHOOD ESTIMATORS.
C
      FNONE=FLOAT(NONE)
      FNTWO=FLOAT(NTWO)
      VMEAN2=V2/FNTWO
      VMEAN1=V1/FNONE
      VVAR1=2.0*FLOAT(NONE-1)*V1**2/(FNONE**2)
      VVAR2=2.0*FLOAT(NTWO-1)*V2**2/(FNTWO**2)
C
C     EVALUATE DERIVATIVES OF X1 AND X2 WITH RESPECT TO U1, U2, V1, V2.
C
      TERM3=1.0/TERM2
      TERM4=SD1SD2*UDIFF*TERM3
      TERM5=(SD1SD2/2.0)*(VDIFF/V1+ALOG(V2/V1))*TERM3
      TERM6=(SD1SD2/2.0)*(VDIFF/V2+ALOG(V2/V1))*TERM3
      DX1U1=(V2-TERM4)/VDIFF
      DX2U1=(V2+TERM4)/VDIFF
      DX1U2=(-V1+TERM4)/VDIFF
      DX2U2=(-V1-TERM4)/VDIFF
      DX1V1=(-U2-SD2*TERM3/(2.0*SD1)+TERM5+X1)/VDIFF
      DX2V1=(-U2+SD2*TERM3/(2.0*SD1)-TERM5+X2)/VDIFF
      DX1V2=(U1-SD1*TERM3/(2.0*SD2)-TERM6-X1)/VDIFF
      DX2V2=(U1+SD1*TERM3/(2.0*SD2)+TERM6-X2)/VDIFF
C
C     CALCULATE THE VARIANCE OF OVLHAT.
C
      PHI11=DNORML(Z11)
      PHI12=DNORML(Z12)
      PHI21=DNORML(Z21)
      PHI22=DNORML(Z22)
      PHI11S=PHI11/SD1
      PHI12S=PHI12/SD2
      PHI21S=PHI21/SD1
      PHI22S=PHI22/SD2
      PTERM1=PHI11S-PHI12S
      PTERM2=PHI22S-PHI21S
      VTERM1=(PTERM1*DX1U1+PTERM2*DX2U1-PHI11S+PHI21S)**2
      VTERM2=(PTERM1*DX1U2+PTERM2*DX2U2-PHI22S+PHI12S)**2
      VTERM3=(PTERM1*DX1V1+PTERM2*DX2V1+(PHI21*Z21-PHI11*Z11)/(2.0*V1))*
     1*2
      VTERM4=(PTERM1*DX1V2+PTERM2*DX2V2+(PHI12*Z12-PHI22*Z22)/(2.0*V2))*
     1*2
      VOVL=VTERM1*VMEAN1+VTERM2*VMEAN2+VTERM3*VVAR1+VTERM4*VVAR2
C
C     PRINT INTERMEDIATE CALCULATIONS AND RESULTS (IF IWRITE=0).
C
      IF (IWRITE .EQ. 0) GO TO 999
      WRITE(6,191) NONE
  191 FORMAT(1H0,'THE SIZE OF THE SAMPLE WITH THE SMALLER VARIANCE =',I5
     1)
      WRITE(6,192) U1
```

```
192   FORMAT(1HO,'ITS MEAN =',F20.8)
      WRITE(6,193) V1
193   FORMAT(1HO,'ITS VARIANCE =',F20.8)
      WRITE(6,194) NTWO
194   FORMAT(1HO,'THE SIZE OF THE SAMPLE WITH THE LARGER VARIANCE =',I5)
      WRITE(6,192) U2
      WRITE(6,193) V2
      WRITE(6,203) X1,X2
203   FORMAT(1HO,'THE CROSSING POINTS =',F20.8,F20.8)
      WRITE(6,205)
205   FORMAT(1HO,'THE DERIVATIVES OF THE CROSSING POINTS')
      WRITE(6,206) DX1U1,DX2U1
206   FORMAT(1HO,'WITH RESPECT TO THE FIRST MEAN =',F20.8,F20.8)
      WRITE(6,207) DX1U2,DX2U2
207   FORMAT(1HO,'WITH RESPECT TO THE SECOND MEAN =',F20.8,F20.8)
      WRITE(6,208) DX1V1,DX2V1
208   FORMAT(1HO,'WITH RESPECT TO THE FIRST VARIANCE =',F20.8,F20.8)
      WRITE(6,209) DX1V2,DX2V2
209   FORMAT(1HO,'WITH RESPECT TO THE SECOND VARIANCE =',F20.8,F20.8)
      WRITE(6,210) Z11
210   FORMAT(1HO,'THE LOWER CROSSING POINT STANDARDIZED TO DISTRIBUTION
     1ONE =',F20.8)
      WRITE(6,215) Z12
215   FORMAT(1HO,'THE LOWER CROSSING POINT STANDARDIZED TO DISTRIBUTION
     1TWO =',F20.8)
      WRITE(6,216) Z21
216   FORMAT(1HO,'THE UPPER CROSSING POINT STANDARDIZED TO DISTRIBUTION
     1ONE =',F20.8)
      WRITE(6,217) Z22
217   FORMAT(1HO,'THE UPPER CROSSING POINT STANDARDIZED TO DISTRIBUTION
     1TWO =',F20.8)
      WRITE(6,218) VMEAN1
218   FORMAT(1HO,'THE VARIANCE OF THE FIRST SAMPLE MEAN =',F20.8)
      WRITE(6,219) VMEAN2
219   FORMAT(1HO,'THE VARIANCE OF THE SECOND SAMPLE MEAN =',F20.8)
      WRITE(6,220) VVAR1
220   FORMAT(1HO,'THE VARIANCE OF THE FIRST SAMPLE VARIANCE =',F20.8)
      WRITE(6,221) VVAR2
221   FORMAT(1HO,'THE VARIANCE OF THE SECOND SAMPLE VARIANCE =',F20.8)
      WRITE(6,222) OVL
222   FORMAT(1HO,'THE OVERLAPPING COEFFICIENT =',F20.8)
      WRITE(6,223) VOVL
223   FORMAT(1HO,'THE VARIANCE OF THE SAMPLE OVERLAPPING COEFFICIENT =',
     1F20.8)
      GO TO 999
997   WRITE(6,998)
998   FORMAT(1HO,'THE VARIANCES ARE EQUAL')
      OVL=0.0
      VOVL=0.0
999   CONTINUE
      RETURN
      END
```

Subroutine PLCDFS

The subroutine PLCDFS plots two empirical distribution functions using the IMSL routine USPDF,

Called subroutines: USPDF (IMSL).

```
      SUBROUTINE PLCDFS(X1,N1,X2,N2,NTOTAL)
C
C     SUBROUTINE PLCDFS CALLS IMSL ROUTINE USPDF() TO OBTAIN A PLOT OF
C     THE TWO EMPIRICAL DISTRIBUTION FUNCTIONS.  X1 IS THE DATA FROM THE
C     FIRST SAMPLE OF SIZE N1.  X2 IS THE DATA FROM THE SECOND SAMPLE OF
C     SIZE N2.  TO MAKE LIFE SIMPLE, THE SUM OF N1+N2=NTOTAL IS ALSO
C     READ INTO THE ROUTINE.
C
      DIMENSION XALL(4000),WHERE(4000,2),IRHERE(4000)
      DIMENSION X1(N1),X2(N2)
      DO 20 I=1,NTOTAL
      IF (I .GT. N1) GO TO 10
      XALL(I)=X1(I)
      GO TO 20
 10   XALL(I)=X2(I-N1)
 20   CONTINUE
      CALL USPDF(XALL,N1,N2,WHERE,4000,IRHERE)
      RETURN
      END
```

Subroutine PRTSPL

The subroutine PRTSPL prints the output of the subroutine BSPLDF.

Called subroutines: USWFV (IMSL).

```
      SUBROUTINE PRTSPL(MODE,RSSQ,NBKPT,BKPT,NCOEFF,COEFF)
C
C     PRINT BSPLDF() OUTPUT.
C
      DIMENSION BKPT(NBKPT),COEFF(NCOEFF)
      WRITE(6,10) MODE
 10   FORMAT(1H0,'THE HANSON DIAGNOSTIC MODE:',I5)
      WRITE(6,20) RSSQ
 20   FORMAT(1H0,'THE RESIDUAL SUM OF SQUARES:',F20.10)
      WRITE(6,30) NBKPT
 30   FORMAT(1H0,'THE NUMBER OF BREAKPOINTS:',I5)
      WRITE(6,40) NCOEFF
 40   FORMAT(1H0,'THE NUMBER OF B-SPLINE COEFFICIENTS:',I5)
      CALL USWFV('BREAKPOINTS',11,BKPT,NBKPT,1,3)
      CALL USWFV('SPLINE COEFFICIENTS',19,COEFF,NCOEFF,1,3)
      RETURN
      END
```

Subroutine RESAMP

The subroutine RESAMP obtains a simple random sample, with replacement, from the original sample for bootstrap replications.

Called subroutines: GGUBS (IMSL).

```
      SUBROUTINE RESAMP(XSEED,NDATA,XDATA,XRESAM)
C
C     GIVEN A SAMPLE OF DATA XDATA OF SIZE NDATA, ROUTINE RESAMP()
C     GENERATES A SIMPLE RANDOM SAMPLE WITH REPLACEMENT OF SIZE NDATA
C     FROM XDATA USING THE IMSL ROUTINE GGUBS() .
C
C     NOTE: XSEED IS A DOUBLE PRECISION SEED FOR THE IMSL ROUTINE
C     GGUBS().  SEE IMSL DOCUMENTATION FOR REQUIREMENTS.
C
      DIMENSION U(2000)
      DIMENSION XDATA(NDATA),XRESAM(NDATA)
      REAL*8 XSEED
C
C     GENERATE THE ARRAY OF UNIFORM (0,1) RANDOM DEVIATES.
C
      CALL GGUBS(XSEED,NDATA,U)
C
C     CONSTRUCT NEW SAMPLE ARRAY.
C
      FN=FLOAT(NDATA)
      DO 10 I=1,NDATA
      ISUB=IFIX(FN*U(I))+1
         XRESAM(I)=XDATA(ISUB)
 10   CONTINUE
      RETURN
      END
```

Subroutine RSSQDF

The purpose of the subroutine RSSQDF is to compute the residual sum of squares for the fitted B-spline estimate of an unknown distribution function.

Called subroutines: BVALUE (de Boor).

```
      SUBROUTINE RSSQDF(NDATA,XDATA,YDATA,NORD,BKPT,NBKPT,COEFF,NCOEFF,R
     1SSQ)
C
C     CALCULATE AND PRINT RESIDUAL SUM OF SQUARES FOR B-SPLINE DF.
C
      DIMENSION XDATA(NDATA),YDATA(NDATA)
      DIMENSION BKPT(40),COEFF(50)
      RSSQ=0.0
      DO 10 I=1,NDATA
      YHAT=BVALUE(BKPT,COEFF,NCOEFF,NORD,XDATA(I),0)
      SQDIFF=(YDATA(I)-YHAT)**2
      RSSQ=RSSQ+SQDIFF
 10   CONTINUE
      RETURN
      END
```

Subroutine SPLOVL

The object of the subroutine SPLOVL is to obtain $\tilde{OVL}$ using quadratic spline estimates of two unknown distribution functions.

Called subroutines: BSPLDF, BVALUE (de Boor), EMPCDF, PLCDFS, PRTSPL, USPLO (IMSL), VSRTA (IMSL).

```
      SUBROUTINE SPLOVL(NONE,XONE,NTWO,XTWO,IPLOT,IWRITE,OVLSPL)
C
C     ROUTINE SPLOVL COMPUTES A SPLINE-FUNCTION ESTIMATE OF OVL.
C     DATA ARE ASSUMED TO BE TRANSFORMED TO THE INTERVAL (0,1).
C
      DIMENSION XONE(2000),XTWO(2000)
      DIMENSION XCDF1(2000),YCDF1(2000),SDCDF1(2000)
      DIMENSION XCDF2(2000),YCDF2(2000),SDCDF2(2000)
      DIMENSION BKPT1(50),BKPT2(50)
      REAL COEFF1(50)/50*0.0/,COEFF2(50)/50*0.0/
      DIMENSION BKPTS(100),UBKPT(100)
      DIMENSION XEST(101),DFEST(101,2),PDFEST(101,2)
      REAL RPLOT0(4)/0.0,0.0,0.0,1.0/,RPLOT1(4)/0.0,0.0,0.0,0.0/
C
C     GET EMPIRICAL DISTRIBUTION FUNCTIONS FOR THE TWO SAMPLES.
C
      CALL EMPCDF(XONE,NONE,NCDF1,XCDF1,YCDF1,SDCDF1)
      CALL EMPCDF(XTWO,NTWO,NCDF2,XCDF2,YCDF2,SDCDF2)
C
C     GET B-SPLINES FOR THE TWO EMIRICAL DISTRIBUTION FUNCTIONS.
C
      NORD=3
      BKLOW=0.0
      BKUP=1.0
      IPASS=2
      CALL BSPLDF(NCDF1,XCDF1,YCDF1,SDCDF1,NORD,BKLOW,BKUP,IPASS,NBKPT1,
     1BKPT1,NCOEF1,COEFF1,RSSQ1,MODE1)
      CALL BSPLDF(NCDF2,XCDF2,YCDF2,SDCDF2,NORD,BKLOW,BKUP,IPASS,NBKPT2,
     1BKPT2,NCOEF2,COEFF2,RSSQ2,MODE2)
C
C     CREATE UNION SET OF THE TWO SETS OF BREAKPOINTS.
C     VECTOR UBKPTS OF LENGTH NUBKPTS CONTAINS THIS UNION WITH ELEMENTS
C     SORTED IN INCREASING MAGNITUDE BY IMSL ROUTINE VSRTA().
C
      NBKPTS=NBKPT1+NBKPT2
      DO 20 I=1,NBKPTS
      IF (I .GT. NBKPT1) GO TO 10
      BKPTS(I)=BKPT1(I)
      GO TO 20
 10   BKPTS(I)=BKPT2(I-NBKPT1)
 20   CONTINUE
C
C     DELETE DUPLICATE BREAKPOINTS AND BREAKPOINTS WHICH MAY HAVE BEEN
C     DEFINED OUTSIDE THE INTERVAL (BKLOW,BKUP).
C
      CALL VSRTA(BKPTS,NBKPTS)
      NUBKPT=0
      DO 40 I=1,NBKPTS
      IF (I .EQ. 1) GO TO 30
      IF (BKPTS(I) .EQ. BKPTS(I-1)) GO TO 40
 30   IF (BKPTS(I) .LT. BKLOW) GO TO 40
      IF (BKPTS(I) .GT. BKUP) GO TO 40
      NUBKPT=NUBKPT+1
      UBKPT(NUBKPT)=BKPTS(I)
 40   CONTINUE
C
C     CALCULATE INTERVAL AREAS AND OVL.
C
C     DIFF IS THE DIFFERENCE IN ESTIMATED DENSITIES (DENSITY TWO MINUS
```

```
C     DENSITY ONE) AT THE BREAKPOINT I;  DIFLST IS THIS DIFFERENCE AT
C     THE BREAKPOINT (I-1).
C     AREA IS THE AREA UNDER MIN(DENSITY ONE, DENSITY TWO) IN INTERVAL
C     BETWEEN UBKPT(I-1) AND UBKPT(I).  THESE AREAS ARE SUMMED TO
C     CALCULATE OVLSPL, THE ESTIMATE OF OVL BASED ON THE B-SPLINE
C     DISTRIBUTION FUNCTIONS AND DENSITIES.
C
C     NOTE:  FUNCTION BVALUE(BKPT,COEFF,NCOEFF,NORD,X,I) EVALUATES THE
C     I-TH DERIVATIVE OF THE B-SPLINE GIVEN BY BKPT,COEFF, NCOEFF, AND
C     NORD AT THE POINT X.  SEE DE BOOR, P. 144.
C
      OVLSPL=0.0
C
C     DIFLST IS INITIALLY SET TO ZERO BECAUSE OF THE CONSTRAINT THAT THE
C     ESTIMATED DENSITIES MUST BE ZERO AT THE BREAKPOINT ZERO.
C
      DIFLST=0.0
      DO 200 I=2,NUBKPT
      J=I-1
      DIFF=BVALUE(BKPT2,COEFF2,NCOEF2,NORD,UBKPT(I),1)-BVALUE(BKPT1,COEF
     1F1,NCOEF1,NORD,UBKPT(I),1)
      IF (DIFF) 110,120,130
  110 IF (DIFLST) 170,170,140
  120 IF (DIFLST) 170,160,160
  130 IF (DIFLST) 150,160,160
  140 XCROSS=UBKPT(J)+DIFLST*(UBKPT(I)-UBKPT(J))/(DIFLST-DIFF)
      AREA=BVALUE(BKPT1,COEFF1,NCOEF1,NORD,XCROSS,0)-BVALUE(BKPT1,COEFF1
     1,NCOEF1,NORD,UBKPT(J),0)+BVALUE(BKPT2,COEFF2,NCOEF2,NORD,UBKPT(I),
     20)-BVALUE(BKPT2,COEFF2,NCOEF2,NORD,XCROSS,0)
      GO TO 180
  150 XCROSS=UBKPT(J)+DIFLST*(UBKPT(I)-UBKPT(J))/(DIFLST-DIFF)
      AREA=BVALUE(BKPT2,COEFF2,NCOEF2,NORD,XCROSS,0)-BVALUE(BKPT2,COEFF2
     1,NCOEF2,NORD,UBKPT(J),0)+BVALUE(BKPT1,COEFF1,NCOEF1,NORD,UBKPT(I),
     20)-BVALUE(BKPT1,COEFF1,NCOEF1,NORD,XCROSS,0)
      GO TO 180
  160 AREA=BVALUE(BKPT1,COEFF1,NCOEF1,NORD,UBKPT(I),0)-BVALUE(BKPT1,COEF
     1F1,NCOEF1,NORD,UBKPT(J),0)
      GO TO 180
  170 AREA=BVALUE(BKPT2,COEFF2,NCOEF2,NORD,UBKPT(I),0)-BVALUE(BKPT2,COEF
     1F2,NCOEF2,NORD,UBKPT(J),0)
  180 OVLSPL=OVLSPL+AREA
      DIFLST=DIFF
  200 CONTINUE
C
C     PRINT THE B-SPLINE ESTIMATE OF OVL, A PLOT OF THE TWO EMPIRICAL
C     DISTRIBUTION FUNCTIONS, AND PLOTS OF THE B-SPLINE DISTRIBUTION
C     FUNCTIONS AND DENSITIES.  (IF IPLOT .NE. 0)
C
      IF (IPLOT .EQ. 0) GO TO 900
      WRITE(6,210) OVLSPL
  210 FORMAT(1HO,'B-SPLINE ESTIMATED OVERLAPPING COEFFICIENT =',F20.10)
C
C     PLOT THE TWO EMPIRICAL DISTRIBUTION FUNCTIONS.
C
      NTOTAL=NONE+NTWO
      CALL PLCDFS(XONE,NONE,XTWO,NTWO,NTOTAL)
C
C     PLOT B-SPLINE ESTIMATED DF AND PDF
C
```

```
      RANGE=BKUP-BKLOW
      DIV=FLOAT(100)
      DO 800 I=1,101
C
C     GENERATE PLOTTING POINTS.
C
      XCAL=FLOAT(I-1)*RANGE/DIV
      XEST(I)=XCAL
C
C     GET B-SPLINE ESTIMATES OF DISTRIBUTION FUNCTIONS.
C
      DFEST(I,1)=BVALUE(BKPT1,COEFF1,NCOEF1,NORD,XCAL,0)
      DFEST(I,2)=BVALUE(BKPT2,COEFF2,NCOEF2,NORD,XCAL,0)
C
C     GET B-SPLINE ESTIMATES OF DENSITY FUNCTIONS.
C
      PDFEST(I,1)=BVALUE(BKPT1,COEFF1,NCOEF1,NORD,XCAL,1)
      PDFEST(I,2)=BVALUE(BKPT2,COEFF2,NCOEF2,NORD,XCAL,1)
 800  CONTINUE
C
C     PLOT ESTIMATED DISTRIBUTION FUNCTIONS.
C
      CALL USPLO(XEST,DFEST,101,101,2,1,'B-SPLINE ESTIMATED DISTRIBUTION
     1 FUNCTIONS',41,'X',1,'ESTIMATED DF AT X',17,RPLOT0,2H12,1,IER0)
C
C     PLOT ESTIMATED DENSITY FUNCTIONS.
C
      CALL USPLO(XEST,PDFEST,101,101,2,1,'B-SPLINE ESTIMATED DENSITY FUN
     1CTIONS',36,'X',1,'ESTIMATED DENSITY AT X',22,RPLOT1,2H12,1,IER1)
C
C     PRINT THE OUTPUT OF THE B-SPLINE ROUTINES USED TO ESTIMATE THE
C      TWO DISTRIBUTION FUNCTIONS.  (IF IWRITE .NE. 0)
C
 900  IF (IWRITE .EQ. 0) GO TO 999
      WRITE(6,910)
 910  FORMAT(1H0,'B-SPLINE RESULTS FOR THE FIRST SAMPLE')
      CALL PRTSPL(MODE1,RSSQ1,NBKPT1,BKPT1,NCOEF1,COEFF1)
      WRITE(6,920)
 920  FORMAT(1H0,'B-SPLINE RESULTS FOR THE SECOND SAMPLE')
      CALL PRTSPL(MODE2,RSSQ2,NBKPT2,BKPT2,NCOEF2,COEFF2)
 999  CONTINUE
      RETURN
      END
```

Subroutine TRANSF

The purpose of the subroutine TRANSF is to apply the selected tranformation to the data in the array X, thus mapping the data to the interval [0,1].

Called subroutines: none.

```
      SUBROUTINE TRANSF(N,X,ITRANS,A,B)
C
C     SUBROUTINE TRANSF() APPLIES THE TRANSFORMATION INDICATED BY THE
C     INPUT VARIABLE ITRANS TO THE DATA IN THE ARRAY X.  N IS THE LENGTH
C     OF X.  ITRANS SET TO ZERO  RETURNS THE UNTRANSFORMED ARRAY X.
C     INPUT VARIABLES A AND B ARE USED FOR TRANSFORMATION ONE ONLY.
C
C          ITRANS=0, THE TRANSFORMATION IS X=X ;
C
C          ITRANS=1, THE TRANSFORMATION IS X=(X-A)/(B-A) ;
C
C          ITRANS=2, THE TRANSFORMATION IS X=X/(1.0+X) ;
C
C          ITRANS=3, THE TRANSFORMATION IS X=EXP(X)/(1.0+EXP(X)) .
C
      DIMENSION X(N)
      IF (ITRANS .EQ. 0) GO TO 400
C
      IF (ITRANS .GT. 1) GO TO 200
C
C     APPLY FIRST TRANSFORMATION.
C
      DIVIDE=B-A
      DO 10 I=1,N
         X(I)=(X(I)-A)/DIVIDE
 10      CONTINUE
      GO TO 400
C
 200  IF (ITRANS .GT. 2) GO TO 300
C
C     APPLY SECOND TRANSFORMATION.
C
      DO 20 I=1,N
         X(I)=X(I)/(1.0+X(I))
 20      CONTINUE
      GO TO 400
C
 300  IF (ITRANS .GT. 3) GO TO 400
C
C     APPLY THIRD TRANSFORMATION.
C
      DO 30 I=1,N
         Y=EXP(X(I))
         X(I)=Y/(1.0+Y)
 30      CONTINUE
C
 400  CONTINUE
      RETURN
      END
```

LIST OF REFERENCES

Anderson, T. W. 1958. An introduction to multivariate statistical analysis. New York: John Wiley.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. 1975. Discrete multivariate analysis: theory and practice. Cambridge: The MIT Press.

Boneva, L. I., Kendall, D., and Stefanov, I. 1971. Spline transformations: three new diagnostic aids for the statistical data analyst. Journal of the Royal Statistical Society, ser. B. 33: 1-37.

Boring, E. G. 1919. Mathematical vs. scientific significance. Psychological Bulletin. 16: 335-38.

Box, G. E. P., and Cox, D. R. 1964. An analysis of transformations. Journal of the Royal Statistical Society, ser. B. 26: 211-52.

Bradley, E. L., and Piantadosi, S. 1982. The overlapping coefficient as a measure of agreement between distributions. Paper presented at the spring meeting of the Alabama chapter, American Statistical Association, Birmingham, Alabama, 27 February 1982.

Buse, A., and Lim, L. 1977. Cubic splines as a special case of restricted least squares. Journal of the American Statistical Association. 72: 64-68.

Chan, T. F., and Lewis, J. G. 1979. Computing standard deviations: accuracy. Communications of the Association for Computational Mathematics. 22: 526-31.

Cheng, R. C. H., and Iles, T. C. 1983. Confidence bands for cumulative distribution functions of continuous random variables. Technometrics. 25: 77-86.

Cohen, J. 1962. The statistical power of abnormal-social psychological research: a review. Journal of Abnormal and Social Psychology. 65: 145-53.

_____. 1977. Statistical power analysis for the behavioral sciences. Rev. ed. New York: Academic Press.

Cohen, J. K., Falk, R. F., and Cortese, C. F. 1976. Reply to Taeuber and Taeuber. American Sociological Review. 41: 889-93.

Cortese, C. F., Falk, R. F., and Cohen, J. K. 1976. Further considerations on the methodological analysis of segregation indices. American Sociological Review. 41: 630-37.

_____. 1978. Understanding the standardized index of dissimilarity: reply to Massey. American Sociological Review. 43: 590-92.

de Boor, C. 1978. A practical guide to splines. Applied mathematical sciences, vol. 27. New York: Springer-Verlag.

de Montricher, G. F., Tapia, R. A., and Thompson, J. R. 1975. Nonparametric maximum likelihood estimation of probability densities by penalty function methods. Annals of Statistics. 3: 1329-48.

Duncan, O. D., and Duncan, B. 1955. A methodological analysis of segregation indices. American Sociological Review. 20: 210-17.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics. 7: 1-26.

_____. 1981. Nonparametric estimates of standard error: the jackknife, the bootstrap, and other methods. Biometrika. 68: 589-99.

_____. 1982. The jackknife, the bootstrap, and other resampling plans. CBMS-NSF regional conference series in applied mathematics, no. 38. Philadelphia: Society for Industrial and Applied Mathematics.

_____. 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American Statistical Association. 73: 316-31.

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

Elandt, R. C. 1961. The folded normal distribution: two methods of estimating parameters from moments. Technometrics. 3: 551-62.

Elgie, R. A. 1979. The segregation of socioeconomic groups in urban areas: a comment. Urban Studies. 16: 191-95.

Falk, R. F., Cortese, C. F., and Cohen, J. K. 1978. Utilizing standardized indices of segregation: comment on Winship. Social Forces. 57: 713-16.

Freeman, G. H., and Halton, J. H. 1951. Note on an exact treatment of contingency, goodness of fit, and other problems of significance. Biometrika. 38: 141-49.

Gastwirth, J. L. 1973. Measurement of earnings differentials between

the sexes. In American Statistical Association: 1973 proceedings of the social statistics section, pp. 133-37. Washington: American Statistical Association.

_____. 1975. Statistical measures of earnings differentials. American Statistician. 29: 32-35.

Gibbons, J. D. 1971. Nonparametric statistical inference. New York: McGraw-Hill.

Goodman, L. A., and Kruskal, W. H. 1979. Measures of association for cross classification. New York: Springer-Verlag.

Hanson, R. J. 1979. Constrained least squares curve fitting to discrete data using B-splines: a user's guide. SAND78-1291. Springfield: U. S. Department of Commerce, Technical Information Service.

Hardison, C. D., Quade, D., and Langston, R. D. 1983. Nine functions for probability distributions. In SUGI supplemental library user's guide, pp. 229-36. 1983 ed. Cary, North Carolina: SAS Institute.

Hornseth, R. A. 1947. A note on "The measurement of ecological segregation" by Julius Jahn, Calvin F. Schmid, and Clarence Schrag. American Sociological Review. 12: 603-4.

Hout, M. 1983. Mobility tables. Quantitative applications in the social sciences, no. 31. Beverly Hills: Sage.

IMSL. 1982. IMSL library reference manual. 9th ed. 4 vols. Houston: IMSL.

Inman, H. F. 1981. Migration in the cotton South: the geographic mobility of Alabama farmers, 1850-1860. M.A. thesis, University of Alabama in Birmingham.

Jahn, J. A. 1950. The measurement of segregation: derivation of an index based on the criterion of reproducibility. American Sociological Review. 15: 100-4.

Jahn, J. A., Schmid, C. F., and Schrag, C. C. 1947. The measurement of ecological segregation. American Sociological Review. 12: 293-303.

_____. 1948. Rejoinder to Dr. Hornseth's note on "The measurement of ecological segregation." American Sociological Review. 13: 216-17.

Johnson, N. L. 1962. The folded normal distribution: accuracy of estimation by maximum likelihood. Technometrics. 4: 249-56.

Johnson, N. L., and Kotz, S. 1969. Discrete distributions. Distribu-

tions in statistics. New York: John Wiley.

_____. 1970. Continuous univariate distributions. Vol. 2. Distributions in statistics. New York: John Wiley.

Johnson, R. A., and Wichern, D. W. 1982. Applied multivariate analysis. Englewood Cliffs: Prentice-Hall.

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

_____. 1979. The advanced theory of statistics. Vol. 2: Inference and relationship. 4th ed. New York: Macmillan.

Kestenbaum, B. 1980. Notes on the index of dissimilarity: a research note. Social Forces. 59: 275-80.

Kozak, J. 1980. On the choice of the exterior knots in the B-spline basis for a spline space. MRC technical summary report, no. 2148. Madison: Mathematics Research Center, University of Wisconsin--Madison.

Krishnaiah, P. R., Hagis, P, and Steinberg, L. 1963. A note on the bivariate chi distribution. SIAM Review. 5: 140-44.

Leone, F. C., Nelson, L. S., and Nottingham, R. B. 1961. The folded normal distribution. Technometrics. 3: 543-50.

Lii, K. S., and Rosenblatt, M. 1975. Asymptotic behavior of a spline estimate of a density function. Computers and Mathematics with Applications. 1: 223-35.

Marx, W. 1976a. Die Messung der Assoziativen Bedeutungsähnlichkeit. Zeitschrift für experimentelle und angewandte Psychologie. 23: 62-76.

_____. 1976b. Die statistische Sicherung des Überlappungs-Koeffizienten. Zeitschrift für experimentelle und angewandte Psychologie. 23: 267-70.

Massey, D. S. 1978. On the measurement of segregation as a random variable. American Sociological Review. 43: 587-90.

Merschrod, K. 1981. The index of dissimilarity as a measure of inequality. Quality and Quantity. 15: 403-11.

Moore, D. S. 1979. Statistics: concepts and controversies. San Francisco: W. H. Freeman.

Morgan, B. S., and Norbury, J. 1981. Some further observations on the index of residential differentiation. Demography. 18: 251-56.

Nabeya, S. 1951. Absolute moments in 2-dimensional normal distribution. Annals of the Institute of Statistical Mathematics (Tokyo). 3: 2-6.

_____. 1952. Absolute moments in 3-dimensional normal distribution. Annals of the Institute of Statistical Mathematics (Tokyo). 4: 15-30.

Pearson, E. S. 1965. Some incidents in the early history of biometry and statistics, 1890-94. Biometrika. 52: 3-18. Reprinted 1970, in Studies in the history of statistics and probability, 1: 323-38. Edited by E. S. Pearson and M. Kendall. London: Charles Griffin.

Rosenblatt, M. 1956. Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics. 27: 832-37.

_____. 1977. Some problems in approximation and estimation. In Proceedings of computer science and statistics: tenth annual conference on the interface, pp. 184-88. National Bureau of Standards special publication, no. 503. Washington: National Bureau of Standards.

Sakoda, J. M. 1981. A generalized index of dissimilarity. Demography. 18: 245-50.

Sampson, P. D. 1979. Comment on "Splines and restricted least squares." Journal of the American Statistical Association. 74: 303-5.

SAS. 1982. SAS user's guide. 1982 ed. 2 vols. Cary, North Carolina: SAS Institute.

Sheehan, T. J. 1980. The medical literature: let the reader beware. Archives of Internal Medicine. 140: 472-74.

Smith, P. L. 1979. Splines as a useful and convenient statistical tool. American Statistician. 33: 57-62.

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: 123-43.

_____. 1979. The sampling distribution of the W statistic of disjunction for the arbitrary division of a random rectangular distribution. Mathematical Geology. 11: 423-42.

Snedecor, G. W., and Cochran, W. G. 1980. Statistical methods. 7th ed. Ames: Iowa State University Press.

Snee, R. D., and Pfeifer, C. G. 1983. Histograms. In Encyclopedia of

of statistical sciences, 3: 635-40. Edited by N. L. Johnson, S. Kotz, and C. B. Read. 8 vols. New York: John Wiley.

Stephans, M. A. 1974. EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association. 69: 730-37.

Steyn, H. S. 1955. On discrete multivariate distribution functions of hypergeometric type. Indagationes Mathematicae. 17: 588-95.

Sturges, H. A. 1926. The choice of a class interval. Journal of the American Statistical Association. 21: 65-66.

Suits, D. B., Mason, A., and Chan, L. 1978. Spline functions fitted by standard regression methods. Review of Economics and Statistics. 60: 132-39.

Taeuber, K. E., and Taeuber, A. F. 1965. Negroes in cities: residential segregation and neighborhood change. Chicago: Aldine.

_____. 1976. A practitioner's perception on the index of dissimilarity. American Sociological Review. 41: 884-89.

Tukey, J. W. 1957. On the comparative anatomy of transformations. Annals of Mathematical Statistics. 28: 602-32.

Wahba, G. 1975. Interpolating spline methods for density estimation: I. Equispaced knots. Annals of Statistics. 3: 30-48.

Wallis, W. A., and Roberts, H. V. 1956. Statistics: a new approach. New York: Free Press.

Waterman, M. S., and Whiteman, D. E. 1978. Estimation of probability densities by empirical density functions. International Journal for Mathematical Education in Science and Technology. 9: 127-37.

Wegman, E. J. 1972. Nonparametric probability density estimation: I. A summary of available methods. Technometrics. 14: 533-46.

_____. 1982. Density estimation. In Encyclopedia of statistical sciences, 2: 309-15. Edited by N. L. Johnson, S. Kotz, and C. B. Read. 8 vols. New York: John Wiley.

Wegman, E. J., and Wright, I. W. 1983. Splines in statistics. Journal of the American Statistical Association. 78: 351-65.

Weitzman, M. S. 1970. Measures of overlap of income distributions of white and negro families in the United States. Technical paper no. 22. Washington: U. S. Department of Commerce, Bureau of the Census.

Williams, J. J. 1948. Another commentary on so-called segregation

indices. American Sociological Review. 13: 298-303.

Winship, C. 1977. A revaluation of indexes of residential segregation. Social Forces. 55: 1058-66.

_____. 1978. The desirability of using the index of dissimilarity or any adjustment of it for measuring segregation: reply to Falk, Cortese, and Cohen. Social Forces. 57: 717-20.

Wold, S. 1974. Spline functions in data analysis. Technometrics. 16: 1-11.