Communications in Statistics - Theory and Methods

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities

Henry F. Inman [a] & Edwin L. Bradley Jr [b]

[a] Dept.of Mathematical Sciences, Virginia Commonwealth University, Richmond, Virginia, 23284

[b] Dept.of Biostatistics and Biomathematics, University of Alabama at Birmingham, Birmingham, Alabama, 35294

Published online: 27 Jun 2007.

To cite this article: Henry F. Inman & Edwin L. Bradley Jr (1989): The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities, Communications in Statistics - Theory and Methods, 18:10, 3851-3874

To link to this article: http://dx.doi.org/10.1080/03610928908830127

COMMUN. STATIST.-THEORY METH., 18(10), 3851-3874 (1989)

THE OVERLAPPING COEFFICIENT AS A MEASURE OF AGREEMENT BETWEEN PROBABILITY DISTRIBUTIONS AND POINT ESTIMATION OF THE OVERLAP OF TWO NORMAL DENSITIES

Henry F. Inman

Dept. of Mathematical Sciences
Virginia Commonwealth University
Richmond, Virginia 23284

Edwin L. Bradley, Jr.

Dept. of Biostatistics and
Biomathematics
University of Alabama at Birmingham
Birmingham, Alabama 35294

Key words and Phrases: *dissimilarity index*; *Hellinger distance*; *Mahalanobis distance*; *maximum-likelihood estimation*; *standard distance*

ABSTRACT

The overlapping coefficient is defined as a measure of the agreement between two probability distributions. Its relationship to the dissimilarity index and its properties are described. An extensive treatment of maximum-likelihood estimation of the overlap between two normal distributions is presented as an example of estimating the overlapping coefficient from sample data.

INTRODUCTION

The overlapping coefficient (OVL) is a measure of agreement or similarity between two probability distributions or two populations represented by such distributions. Let $f_1(\underline{X})$ and $f_2(\underline{X})$ be probability (density) functions defined on R_n. In the continuous case OVL is formally defined as

$$\text{OVL} = \int_{R_n} \min\,[f_1(\underline{X}),\, f_2(\underline{X})]\, d\underline{X}. \tag{1}$$

In the discrete case OVL can be defined in an analogous form:

$$\text{OVL} = \sum_{\underline{X}} \min[f_1(\underline{X}), f_2(\underline{X})]\,.$$

The depiction of OVL for a simple univariate example in Figure 1 displays the natural interpretation of OVL as a measure of agreement, some obvious properties of this measure, and how OVL is determined when $f_1(\underline{X})$ and $f_2(\underline{X})$ are completely specified. Here it is apparent that OVL is simply the fraction of probability mass under either distribution also common to the other, represented by the shaded area in Figure 1. It immediately follows that the numerical value of OVL is bounded by zero and unity. Furthermore, OVL=0 if and only if $f_1(\underline{X})$ and $f_2(\underline{X})$ are disjoint, and OVL=1 if and only if $f_1(\underline{X})=f_2(\underline{X})$. To compute OVL for two known distributions, one simply determines the points where $f_1(\underline{X})$ and $f_2(\underline{X})$ cross and then evaluates the integral that defines OVL (1). Note that $f_1(\underline{X})$ and $f_2(\underline{X})$ need not be of the same parametric form. Examples of such computations can be found in Bradley and Piantadosi (1982) and Inman (1984).

In a review of measures of similarity, Gower (1985) observes that the overlapping coefficient conforms to a general notion of assessing the similarity of two populations, has a very general form and thus is potentially widely applicable, but remains little used in practice. Like the Kullback discrimination information (Kullback, 1983), OVL is a similarity measure with metric properties. Unlike the Kullback information, OVL possesses no major philosophical foundation. Instead, the chief appeal of OVL is its simplicity and "naturalness" as a measure of agreement between $f_1(\underline{X})$ and $f_2(\underline{X})$. In the context of classifying individuals into one of the two populations represented by these densities on the basis of the observed value $\underline{x}$, OVL is the sum of the conditional misclassification probabilities when the classification rule (assuming continuity) is to assign an individual to the first population if $f_1(\underline{x}) > f_2(\underline{x})$ and to the second population if $f_1(\underline{x}) < f_2(\underline{x})$.

As a measure of agreement, OVL possesses a particularly interesting property: OVL is invariant when a suitable common transformation is made to both distributions. By the usual change of variable in calculus, it is clear that OVL written as

$$\text{OVL} = \int_{R_n} \min \left[f_1(g(\underline{X})) , f_2(g(\underline{X})) \right] \mid dg(\underline{X}) \mid$$

is equal to the original expression for OVL (1) when $g(\underline{X})$ is a continuous differentiable function defined on R_n that is one-to-one and preserves order. For example, the normal theory results discussed below can be extended to situations where a normalizing transformation, like Tukey (1957) or Box and Cox (1964), can be employed.

The complement of OVL is the dissimilarity index D, where D can be defined as

$$D = \tfrac{1}{2} \int_{R_n} |f_1(\underline{X}) - f_2(\underline{X})| \, d\underline{X}$$

in the continuous case, and as

$$D = \tfrac{1}{2} \sum_{\underline{X}} |f_1(\underline{X}) - f_2(\underline{X})|$$

in the discrete case. Since $f_1(\underline{X})$ and $f_2(\underline{X})$ are nonnegative, we can write

$$\min \left[f_1(\underline{X}) , f_2(\underline{X}) \right] = \tfrac{1}{2} \left[f_1(\underline{X}) + f_2(\underline{X}) - | f_1(\underline{X}) - f_2(\underline{X}) | \right]$$

and from (1) demonstrate that OVL=1−D. Thus D represents the fraction of probability mass under either $f_1(\underline{X})$ or $f_2(\underline{X})$ not shared with the other, or the amount of probability mass under one density that must be shifted to obtain the other.

While in more general form D has proved useful in asymptotic estimation theory (for example in the construction of consistent estimators), our interest in OVL has a different emphasis. When working with large samples, the increased power of common statistical procedures permits us to determine that small differences between two populations represented by $f_1(\underline{X})$ and $f_2(\underline{X})$ exist. In such circumstances, a measure of the similarity of the two distributions like OVL seems to us more useful than the simple conclusion that $f_1(\underline{X}) \neq f_2(\underline{X})$ at some level of statistical significance. More generally, the computation of OVL whenever sample data indicate two populations differ serves as an assessment of the meaningfulness of the detected difference. Of course, the distinction between a practical, meaningful, or clinical difference and some statistically significant difference is not new; see Boring (1919), for example. Here we simply suggest that the overlapping coefficient provides one approach to this issue.

In a series of papers, culminating in Bradley (1985), OVL was proposed as a generalized measure of agreement and given its current formulation. However, both OVL and D (or their equivalents) had already been used; in fact, it is as D that this measure first appeared. According to Pearson (1895), he and Yule used D, ignoring the multiplicative constant, as a measure of agreement between various distributions when fitting members of the Pearson system of distributions to sample data. Pearson called this quantity the "areal deviations" of two curves. Thus D antedates Pearson's development of the chi-square method for examining the goodness-of-fit of his distribu-

tions to sample data. The dissimilarity index later became the focus of a somewhat confusing literature as a measure of racial segregation; D was used to compare the relative frequency distributions of white and black residents in subdivisions of geographic units. For an introduction to this literature, see the citations in Duncan and Duncan (1955), Cortese et al. (1976), and Winship (1978). The dissimilarity index D is a special case of the Hellinger or Matusita distance between the densities $f_1(\underline{X})$ and $f_2(\underline{X})$; see Ibragimov and Has'minskii (1981) and the citations in Ahmad (1985). Kamps (1988) notes the relationship among the numerical values of D and several other such measures in a class of probability densities that includes the gamma and Weibull distributions. Weitzman (1970) evidently was the first to use OVL explicitly. He adopted OVL, in its discrete formulation, to compare the income distributions of whites and blacks in the United States; these comparisons provoked a brief critique by Gastwirth (1975).

Recent published work primarily addresses OVL in the special case where $f_1(\underline{X})$ and $f_2(\underline{X})$ are univariate normal densities and the overlapping coefficient must be estimated from sample data. The focus of these investigations is the problem of testing hypotheses concerning the true value of OVL, and point estimation and the properties of point estimators for OVL are either developed incorrectly (Marx, 1976a, 1976b) or taken for granted and ignored (Sneath, 1977, 1979; Mishra et al., 1986). In the discussion that follows, we examine this problem as an illustration of the maximum-likelihood estimation of OVL in a relatively simple distributional setting.

THE OVERLAP OF TWO NORMAL DENSITIES

The computation or estimation of OVL for two normal distributions, with density functions $f_1(X;\mu_1,\sigma_1^2)$ and $f_2(X;\mu_2,\sigma_2^2)$, depends on whether the two variances are equal. Since the estimation of OVL when $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is simpler, has attracted more attention, and leads to firmer conclusions, we consider this situation in some detail. On the other hand, we shall simply summarize the results we have obtained in the more complicated circumstance when $\sigma_1^2 \neq \sigma_2^2$.

The overlap between two normal distributions with equal variances is shown in Figure 1. When $\sigma_1^2 = \sigma_2^2$, the normal density functions intersect at the single value of X equal to $(\mu_1+\mu_2)/2$. Using equation 1 and representing the standard normal distribution function by $\Phi(\cdot)$, we obtain

$$\text{OVL} = 2\,\Phi\left(-\frac{|\mu_1-\mu_2|}{2\sigma}\right) = 2\,\Phi(-\tfrac{1}{2}|\delta|), \tag{2}$$

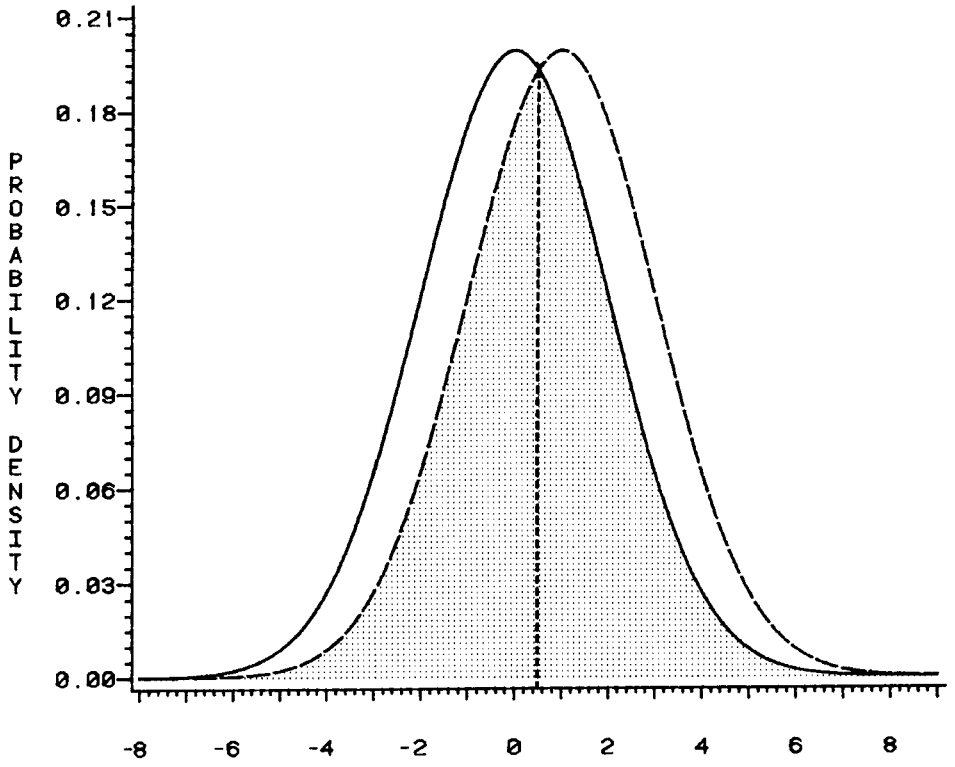

FIG. 1. The overlap between two normal distributions with equal variances. The point of intersection is indicated by the vertical broken line.

where $\delta = (\mu_1 - \mu_2)/\sigma$. In Figure 1, $\mu_1 = 0$, $\mu_2 = 1$, $\sigma^2 = 4$, and OVL=0.80258. If we have two independent simple random samples of sizes n_1 and n_2 from $f_1(X;\mu_1, \sigma^2)$ and $f_2(X;\mu_2,\sigma^2)$ respectively, the invariance property of maximum-likelihood estimators means that the maximum-likelihood estimator for OVL is just

$$\hat{OVL} = 2\Phi\left(-\frac{|\bar{X}_1 - \bar{X}_2|}{2S}\right), \tag{3}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the usual sample means and S is the square-root of the pooled maximum-likelihood estimator for σ^2,

$$S^2 = \frac{\sum_{j=1}^{n_1}(X_{1j} - \bar{X}_1)^2 + \sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2}{n_1 + n_2}.$$

These expressions for OVL and $\hat{OVL}$ demonstrate that the overlapping coefficient in this distributional setting is related to several recognizable quantities. First, OVL and $\hat{OVL}$ are functions of, respectively, the Mahalanobis distance δ^2 and its maximum-likelihood estimator. OVL can also be viewed as a transformation applied to the Kullback discrimination information, which in this circumstance is half the Mahalanobis distance (Kullback, 1983). In the classification problem involving two normal populations, the conditional probabilities of misclassification are equal, so OVL represents twice the misclassification probability; see, for example, Anderson (1984, p. 207). Although the expression for OVL resembles the probability that one independent normal random variable is greater than another (Reiser and Guttman, 1986), the two measures do not appear to be related. Flury and Riedwyl (1986) refer to the quantity $|\bar{X}_1-\bar{X}_2|/S$ as the sample standard distance. (Following their lead, we shall refer to the parameter $|\delta|$ as the standard distance between populations.)

Distribution Function and Moments of $\hat{OVL}$

The sampling distribution of $\hat{OVL}$ can be related directly to the noncentral F distribution. For $0<p<1$, the cumulative distribution function for $\hat{OVL}$ is given by

$$F(p) = P(\hat{OVL} \le p) = P\left[-\frac{|\bar{X}_1 - \bar{X}_2|}{2S} \le \Phi^{-1}\left(\frac{p}{2}\right)\right]$$

$$= P\left\{\left(\frac{\bar{X}_1-\bar{X}_2}{S}\right)^2 \ge 4\left[\Phi^{-1}\left(\frac{p}{2}\right)\right]^2\right\}$$

$$= P\left\{F \ge \frac{4n_1n_2(n_1+n_2-2)}{(n_1+n_2)^2}\left[\Phi^{-1}\left(\frac{p}{2}\right)\right]^2\right\} \qquad (4)$$

where F has a noncentral F distribution with 1 numerator and n_1+n_2-2 denominator degrees of freedom, and noncentrality parameter λ:

$$\lambda = \frac{n_1n_2}{n_1+n_2}\left(\frac{\mu_1-\mu_2}{\sigma}\right)^2 = \frac{n_1n_2}{n_1+n_2}\delta^2.$$

Note that though this sampling distribution is written in terms of the noncentral F, this relationship can also be restated in terms of a folded noncentral t distribution (the positive square-root of the noncentral F). Let t designate the noncentral t random

variable with $n_1 + n_2 - 2$ degrees of freedom and noncentrality parameter $^{+}\sqrt{\lambda}$. Then for $0<p<1$, the distribution function (4) is equivalent to

$$F(p) = 1 - P\left\{ \frac{2\sqrt{n_1n_2(n_1+n_2-2)}}{n_1+n_2}\,\Phi^{-1}\left(\frac{p}{2}\right)<t< \right.$$

$$\left. \frac{-2\sqrt{n_1n_2(n_1+n_2-2)}}{n_1+n_2}\Phi^{-1}\left(\frac{p}{2}\right) \right\}. \tag{5}$$

With this formulation, an approximation to the distribution function of $\hat{OVL}$ can be obtained via the normal approximation to the noncentral t probability required (Abramowitz and Stegun, 1972). Other approximations are possible using distributions asymptotically equivalent to the noncentral F and folded noncentral t, the noncentral χ^2 and folded normal (or noncentral χ) distributions respectively.

Moments (about zero) of the sampling distribution of $\hat{OVL}$ can be calculated numerically from the distribution function, using integration by parts to obtain the following result:

$$\mu_r' = 1 - r\int_0^1 p^{r-1}\,F(p)\,dp\,. \tag{6}$$

for $r=1,2,\cdots$. Examples of such calculations to find the mean and variance of $\hat{OVL}$ are presented and discussed below.

Approximations to the Mean and Variance and the Bias of $\hat{OVL}$

Approximations to the mean and variance of $\hat{OVL}$ can be obtained in a straightforward manner. From equation 3, we see that $\hat{OVL}$ can be written as a function of two independent random variables, $|\bar{X}_1-\bar{X}_2|$ and S^2, which of course have known sampling properties. In particular, because $\bar{X}_1-\bar{X}_2$ follows a known normal distribution, the variate $|\bar{X}_1-\bar{X}_2|$ follows the folded normal distribution (Leone et al., 1961) with mean

$$E(|\bar{X}_1-\bar{X}_2|) = \left[\frac{2(n_1+n_2)}{n_1n_2\pi}\right]^{\frac{1}{2}}\sigma\,\exp\left[\frac{-n_1n_2}{2(n_1+n_2)}\left(\frac{\mu_1-\mu_2}{\sigma}\right)^2\right]$$

$$+\ (\mu_1-\mu_2)\left\{1-2\Phi\left[-\left(\frac{n_1n_2}{n_1+n_2}\right)^{\frac{1}{2}}\ \frac{\mu_1-\mu_2}{\sigma}\right]\right\} \tag{7}$$

and variance

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

Furthermore, based on the relationship between S^2 and the chi-square distribution,

$$\mathrm{E}(S^2) = \frac{(n_1+n_2-2)}{(n_1+n_2)}\,\sigma^2$$

and

$$\mathrm{Var}(S^2) = \frac{2(n_1+n_2-2)}{(n_1+n_2)^2}\,\sigma^4 .$$

Using a first-order Taylor series approximation about $|\mu_1-\mu_2|$ and σ^2 for $\hat{\mathrm{OVL}}$ in equation 3, we can write

$$\hat{\mathrm{OVL}} \cong \mathrm{OVL} - \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\tfrac{1}{8}\left(\tfrac{\mu_1-\mu_2}{\sigma}\right)^2\right]\left[|\bar{X}_1-\bar{X}_2|-|\mu_1-\mu_2| - \frac{|\mu_1-\mu_2|}{2\sigma^2}\,(S^2-\sigma^2)\right].$$

Thus the mean of $\hat{\mathrm{OVL}}$ is approximated by

$$\mathrm{E}(\hat{\mathrm{OVL}}) \cong \mathrm{OVL} - \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\tfrac{1}{8}\left(\tfrac{\mu_1-\mu_2}{\sigma}\right)^2\right]$$

$$\times \left[\mathrm{E}(|\bar{X}_1-\bar{X}_2|)-|\mu_1-\mu_2|\left(1-\frac{1}{n_1+n_2}\right)\right], \tag{8}$$

substituting the expression for $\mathrm{E}(|\bar{X}_1-\bar{X}_2|)$ given in (7). Similarly, the approximate variance of $\hat{\mathrm{OVL}}$ can be obtained from the linear approximation to $\hat{\mathrm{OVL}}$ (the method of statistical differentials, Kendall and Stuart [1977]). We have

$$\mathrm{Var}(\hat{\mathrm{OVL}}) \cong \frac{1}{2\pi} \exp\left[-\frac{1}{4}\left(\frac{\mu_1-\mu_2}{\sigma}\right)^2\right]\left\{\frac{n_1+n_2}{n_1 n_2}\right.$$

$$\left. + \left(\frac{\mu_1-\mu_2}{\sigma}\right)^2\left[1+\frac{(n_1+n_2-2)}{2(n_1+n_2)^2}\right] - \left[\frac{\mathrm{E}(|\bar{X}_1-\bar{X}_2|)}{\sigma}\right]^2\right\}. \tag{9}$$

In passing, we note that when $\mu_1 = \mu_2$, or $\mathrm{OVL}=1$, this approximate variance simplifies to

$$\mathrm{Var}(\hat{\mathrm{OVL}}) \cong \tfrac{1}{2}\left(\tfrac{n_1+n_2}{n_1 n_2}\right)\frac{\pi-2}{\pi^2},$$

which is simply $\mathrm{Var}(|\bar{X}_1-\bar{X}_2|)$ multiplied by the constant $(2\pi\sigma^2)^{-1}$. Although we have not done so here, substitution of (7) into equations 8 and 9 demonstrates that the approximate mean and variance of $\hat{OVL}$ are functions of μ_1, μ_2 and σ^2 only through the parameter δ.

To assess the validity of the approximation formulae, we calculate in Table I the relative errors of the approximations for the indicated values of $n_1=n_2$ and $|\delta|$. Here we present numerical values for the expectation and variance of $\hat{OVL}$ obtained from the first and second noncentral moments computed from (6) and the values obtained from the approximations given by (8) and (9). All computations were performed in double-precision using standard IMSL subroutines. The numerical evaluation of the integrals defined in equation 6, for example, in based on the IMSL adaptive integration algorithm determined by 21-point Gauss-Kronrod quadrature over subintervals defined (in this case) between zero and one.

Inspection of the relative errors in Table I leads to several obvious conclusions. First, the approximation to the expectation of $\hat{OVL}$ consistently overstates the true mean. The relative error of this approximation decreases with increasing sample sizes, but it increases with $|\delta|$. Thus the approximation to the mean of $\hat{OVL}$ performs most poorly when n_1 and n_2 are small and $|\delta|$ is large. However, even when sample sizes are small ($n_1=n_2=25$) the relative error is less than 1.5 percent when $|\delta|$ is less than one. For moderate sample sizes ($n_1=n_2=50$) the relative error of the approximation to the mean barely exceeds 2.0 percent at $|\delta|=3.00$, and when sample sizes become larger (100 or greater) the maximum relative error slightly exceeds 1.0 percent at the same value of $|\delta|$. Second, the approximation to the variance of $\hat{OVL}$ generally understates the true variance of interest. This negative relative error decreases as $|\delta|$ increases, until at $|\delta|\geq 2.00$ we see that the approximation to the variance begins to overstate the true variance. Once again, the greatest relative error in this approximation occurs when n_1 and n_2 are small, while for the larger sample sizes ($n_1=n_2=100$ and $n_1=n_2=200$) the magnitude of the relative error of the variance approximation never exceeds 3.0 percent for all values of $|\delta|$ between zero and three.

The bias of $\hat{OVL}$ as a point estimator for OVL is evident both in the approximation formula for the expectation (8) and in the calculations based on (6) and (8) in Table I. The approximation to the mean in equation 8 indicates that the bias of $\hat{OVL}$ is primarily determined by the bias of $|\bar{X}_1-\bar{X}_2|$ as an estimator for $|\mu_1-\mu_2|$, thus producing the bias of the sample standard distance as an estimator for $|\delta|$. Since

TABLE I

APPROXIMATE AND COMPUTED MEANS AND VARIANCES FOR $\hat{OVL}$

$n_1=n_2$	Mean		Relative	Variance		Relative
	Approximate	Computed	Error(%)	Approximate	Computed	Error(%)
			$\|\delta\|$=0.00, OVL=1.000000			
25	0.909968	0.907326	0.291	0.0046267008	0.0049450734	-6.438
50	0.936338	0.935420	0.981	0.0023133504	0.0023912376	-3.257
100	0.954984	0.954663	0.034	0.0011566752	0.0011759309	-1.637
200	0.968169	0.968056	0.012	0.0005783376	0.0005831244	-0.821
			$\|\delta\|$=0.25, OVL=0.900524			
25	0.875384	0.872597	0.319	0.0075086990	0.0080095090	-6.253
50	0.891524	0.890270	0.141	0.0046660639	0.0048161195	-3.116
100	0.898296	0.897680	0.069	0.0028120632	0.0028555658	-1.523
200	0.900117	0.899808	0.034	0.0015476273	0.0015592059	-0.743
			$\|\delta\|$=0.50, OVL=0.802587			
25	0.795336	0.790340	0.632	0.0109994765	0.0116542414	-5.618
50	0.800344	0.797890	0.308	0.0060436885	0.0062159275	-2.771
100	0.801615	0.800403	0.151	0.0030806095	0.0031236787	-1.379
200	0.802104	0.801503	0.075	0.0015416097	0.0015523322	-0.691
			$\|\delta\|$=0.75, OVL=0.707660			
25	0.701822	0.694661	1.031	0.0116631758	0.0122534972	-4.818
50	0.704868	0.701380	0.497	0.0059104274	0.0060578696	-2.434
100	0.706266	0.704547	0.244	0.0029580300	0.0029947719	-1.227
200	0.706963	0.706110	0.121	0.0014795014	0.0014886710	-0.616
			$\|\delta\|$=1.00, OVL=0.617075			
25	0.610024	0.601232	1.462	0.0110988035	0.0115451646	-3.866
50	0.613554	0.609265	0.704	0.0055653494	0.0056783756	-1.990
100	0.615315	0.613197	0.345	0.0027857761	0.0028141844	-1.009
200	0.616195	0.615143	0.171	0.0013936627	0.0014007817	-0.508
			$\|\delta\|$=1.50, OVL=0.453255			
25	0.444221	0.433902	2.378	0.0092134678	0.0093225263	-1.170
50	0.448738	0.443650	1.147	0.0046271385	0.0046590311	-0.685
100	0.450996	0.448471	0.563	0.0023186702	0.0023272049	-0.367
200	0.452125	0.450868	0.279	0.0011606103	0.0011628133	-0.189
			$\|\delta\|$=2.00, OVL=0.317311			
25	0.307632	0.297956	3.247	0.0069323001	0.0067567611	2.598
50	0.312471	0.307629	1.574	0.0034895700	0.0034499564	1.148
100	0.314891	0.312470	0.775	0.0017506400	0.0017412907	0.537
200	0.316101	0.314890	0.384	0.0008767837	0.0008745164	0.259
			$\|\delta\|$=3.00, OVL=0.133614			
25	0.125843	0.120953	4.043	0.0027913280	0.0024961845	11.824
50	0.129729	0.127171	2.011	0.0014107613	0.0013354025	5.643
100	0.131672	0.130364	1.003	0.0007091550	0.0006901388	2.755
200	0.132643	0.131982	0.501	0.0003555211	0.0003507254	1.367

$E(|\bar{X}_1-\bar{X}_2|) \geq |\mu_1-\mu_2|$, we see from (8) that $E(\hat{OVL}) \leq OVL$, and that the bias of $\hat{OVL}$ decreases with increasing n_1 and n_2, as the bias of $|\bar{X}_1-\bar{X}_2|$ declines. For the values of $|\delta|$ of practical interest, like those in Table I, the tendency of $\hat{OVL}$ to underestimate the true overlap is very obvious. Although this bias decreases as the standard distance increases for all values of $n_1=n_2$, at large values of $|\delta|$ the difference between $E(\hat{OVL})$ and OVL remains noticeable, even for the larger sample sizes.

An alternative approximation to the variance of $\hat{OVL}$ based on the noncentral F random variable in (4) is possible, but we find that such an approximation does not perform as well as the approximation given here. Of course, the results presented as equation 8 and equation 9 permit us to compute approximations to the mean and variance of the maximum-likelihood estimator for the dissimilarity index D, as $\hat{D}=1-\hat{OVL}$, $E(\hat{D})=1-E(\hat{OVL})$, and $Var(\hat{D})=Var(\hat{OVL})$. In addition to the insight the approximations (8) and (9) provide to the sampling behavior of $\hat{OVL}$, the approximation to the variance (9) permits us to obtain an estimate for the true variance of $\hat{OVL}$ from the sample data, substituting the sample estimates for the parameters μ_1, μ_2 and σ^2 into equations 7 and 9 to compute the sample-estimated variance of $\hat{OVL}$. We briefly examine the performance of this use of the variance approximation below.

Asymptotic Behavior of $\hat{OVL}$

As the maximum-likelihood estimator for OVL, $\hat{OVL}$ possesses known asymptotic properties: consistency, unbiasedness, efficiency, and normality. Maximum-likelihood theory, of course, does not guarantee that $\hat{OVL}$ exhibits these properties when n_1 and n_2 are small nor that these limiting properties are attained rapidly as sample sizes increase. To the extent that the linear approximation to $\hat{OVL}$ in terms of $|\bar{X}_1-\bar{X}_2|$ and S^2 holds, we can make the following arguments.

For sufficiently large n_1 and n_2, the distribution of S^2 converges to the normal distribution (Johnson and Kotz, 1970, p. 170).

When $|\delta|>0$, the folded normal distribution of $|\bar{X}_1-\bar{X}_2|$ converges to the normal distribution for sufficiently large n_1 and n_2 (Elandt, 1961, and Leone et al., 1961). But when $\delta=0$ the distribution of $|\bar{X}_1-\bar{X}_2|$ is the half-normal distribution for all finite values of n_1 and n_2.

Thus the sampling distribution of $\hat{OVL}$ displays approximate normality when the sample sizes n_1 and n_2 are sufficiently large only if $|\delta|$ is positive but not large enough to

invalidate the linear representation of $\hat{OVL}$. (Here sufficiently large n_1 and n_2 obviously depends on the magnitude of $|\delta|$, and we shall see below that the upper limit for $|\delta|$ increases with n_1 and n_2.) However, when $\delta=0$ the sampling distribution of $\hat{OVL}$ does not attain even approximate normality for finite sample sizes. These conclusions are illustrated in Figure 2, where for selected values of $n_1 = n_2$ and $|\delta|$ we superimpose the normal density with mean and variance determined by equation 6 over the density of $\hat{OVL}$ obtained by finite-difference approximations to the derivative of the distribution function (5).

The convergence of the sampling distribution of $\hat{OVL}$ to approximate normality can be demonstrated explicitly through direct comparison of the sampling distribution of $\hat{OVL}$ to the normal distribution. Here we shall compare the distribution of $\hat{OVL}$ and the normal distribution by computing the overlap between the theoretical density of $\hat{OVL}$ and the normal densities specified by the mean and variance based on (6) and by the approximate mean and variance based on (8) and (9). The reader will observe that we are using equation 1 to obtain this overlap numerically in an instance where one of the densities involved, that for $\hat{OVL}$, cannot be written in closed form. We have used finite-difference approximations to this density (based on equation 5) to determine where the two densities cross. As above, these computations were performed in double precision using standard IMSL subroutines. The comparison of the sampling distribution of $\hat{OVL}$ to the normal distribution is presented in Table II. The general pattern clearly follows the restrictions on the normality of $\hat{OVL}$ stated above, whether one looks at the overlap of the sampling density of $\hat{OVL}$ and the normal density with the numerically computed moments or the overlap based on the normal density using the approximate moments calculated from equations 8 and 9. For fixed $n_1=n_2$, the distribution of $\hat{OVL}$ becomes increasingly normal as $|\delta|$ increases from zero but then eventually departs from normality as $|\delta|$ becomes large. The effect of large sample sizes is also clear in Table II, as the range of values of $|\delta|$ for which the sampling distribution of $\hat{OVL}$ is approximately normal widens with increasing $n_1=n_2$.

Thus we see that the sampling distribution of $\hat{OVL}$ in the situation of independent simple random samples from two normal distributions with equal variances can be treated as normal, qualified by the constraints on sample sizes and standard distance noted above. Since the conditions for approximate normality of $\hat{OVL}$ are more restrictive than the circumstances in which the approximations to the mean and variance of $\hat{OVL}$ in equations 7 and 8 hold, it appears that when approximate normality prevails,

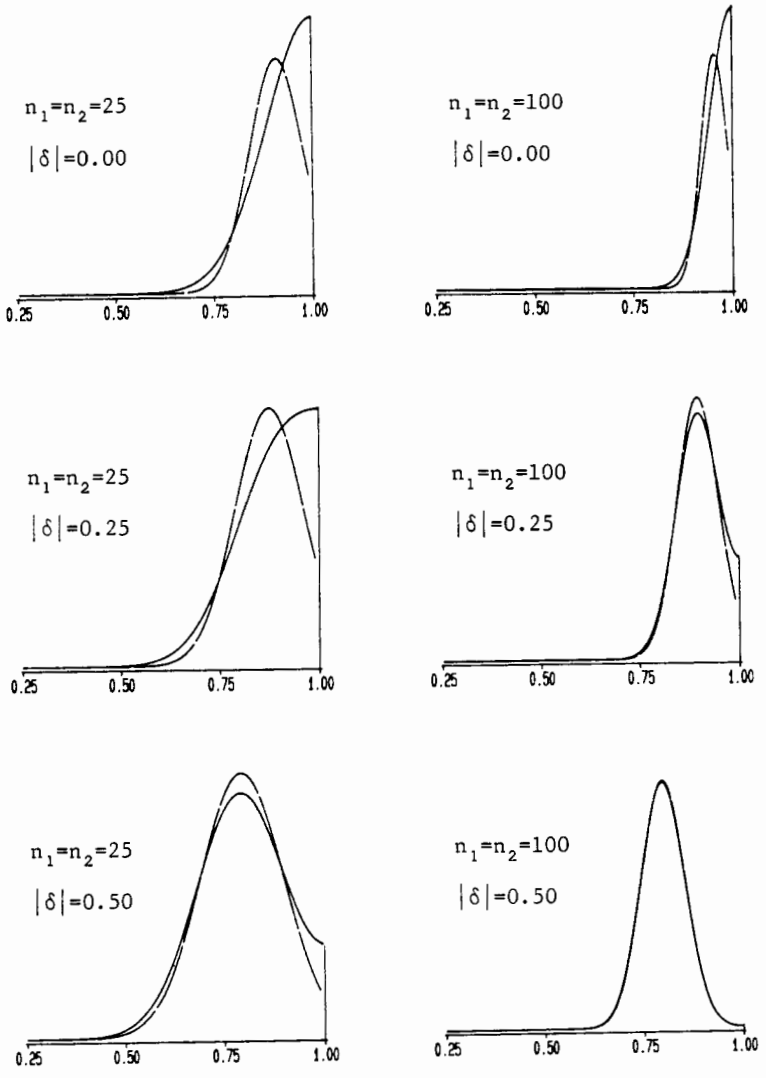

FIG. 2. Finite difference approximations to the sampling density of $\hat{OVL}$ for selected values of $n_1=n_2$ and $|\delta|$. In each panel a portion of the normal density with mean and variance determined from equation 6 is depicted by the broken line. Note that heights and areas are not comparable in different panels.

TABLE II

OVERLAP BETWEEN THE SAMPLING DISTRIBUTION OF $\hat{OVL}$ AND THE NORMAL DENSITY WITH COMPUTED MEAN AND VARIANCE AND APPROXIMATE MEAN AND VARIANCE

$n_1=n_2$	Using Computed Moments	Using Approximate Moments	$n_1=n_2$	Using Computed Moments	Using Approximate Moments
$\lvert\delta\rvert$ = 0.00, OVL=1.000000			$\lvert\delta\rvert$ = 1.00, OVL=0.617075		
25	0.8004859	0.8042282	25	0.9946844	0.9631364
50	0.8012992	0.8033124	50	0.9969276	0.9751919
100	0.8017076	0.8027491	100	0.9980185	0.9828618
200	0.8019118	0.8024411	200	0.9986622	0.9880213
$\lvert\delta\rvert$ = 0.25, OVL=0.900524			$\lvert\delta\rvert$ = 1.50, OVL=0.453255		
25	0.8349188	0.8345348	25	0.9851465	0.9488683
50	0.8805599	0.8791428	50	0.9908519	0.9651485
100	0.9445422	0.9428836	100	0.9939682	0.9757923
200	0.9895731	0.9878918	200	0.9958822	0.9830327
$\lvert\delta\rvert$ = 0.50, OVL=0.802587			$\lvert\delta\rvert$ = 2.00, OVL=0.317311		
25	0.9481384	0.9383492	25	0.9695655	0.9342362
50	0.9905667	0.9801515	50	0.9807906	0.9553635
100	0.9994457	0.9906304	100	0.9871731	0.9690461
200	0.9997318	0.9936800	200	0.9911872	0.9783196
$\lvert\delta\rvert$ = 0.75, OVL=0.707660			$\lvert\delta\rvert$ = 3.00, OVL=0.133614		
25	0.9940877	0.9681107	25	0.9377202	0.9111678
50	0.9984972	0.9805878	50	0.9485196	0.9302285
100	0.9990905	0.9867511	100	0.9651760	0.9521399
200	0.9994002	0.9907789	200	0.9759135	0.9666533

Note: All overlapping coefficients are computed using finite-difference approximations to the derivative of the theoretical distribution function (5).

particularly when sample sizes are large, we may choose to treat $\hat{OVL}$ as normally distributed with mean and variance determined by the approximation formulae. Naturally, these remarks also apply to $\hat{D}=1-\hat{OVL}$ in this distributional setting. (Elsewhere we make the argument for a folded-normal approximation to the sampling distribution of $\hat{D}$, which incorporates the approximate normality we observe here and applies when the standard distance is zero or close to zero; see Inman, 1984.) In passing, we note that an appealing candidate, the standard beta distribution, fails as an approximation to the sampling distribution of $\hat{OVL}$ or $\hat{D}$; this result actually follows from the true sampling distribution of $\hat{OVL}$ given in (4) and the relationship between the noncentral F and the noncentral beta distributions (Johnson and Kotz, 1970, p. 197). We remind the reader that the true sampling distribution for $\hat{OVL}$ is the distribution determined by equations 4 or 5, and tests of hypotheses and confidence intervals for OVL

based on either 4 or 5 will generally prove superior to statistical tests and interval estimators based on the approximate normality of $\hat{OVL}$.

The Variance Approximation as a Point Estimator for Var($\hat{OVL}$)

Ordinarily, the parameters μ_1, μ_2 and σ^2 are unknown, and Var($\hat{OVL}$) would itself be estimated from the sample data. In this context, then, the performance of the approximation (9), with the values of $\bar{X}_1$, $\bar{X}_2$, and S^2 substituted for μ_1, μ_2, and σ^2, as a point estimator for the true variance becomes a question of some interest. To address this issue, we present in Table III a summary of a simple Monte Carlo examination of the sampling behavior of what, for clarity, we shall call the sample-estimated variance of $\hat{OVL}$. For each combination of $|\delta|$ and $n_1=n_2$ in Table III, 5,000 realizations of the sample-estimated variance were obtained using the IMSL double-precision normal and gamma pseudorandom number generators to produce realizations of $\bar{X}_1-\bar{X}_2$ and S^2. We use the mean and standard deviation of these Monte Carlo realizations of the sample-estimated variance to assess the performance of the sample-estimated variance, relative to the actual variance of $\hat{OVL}$ computed numerically from (6), at each combination of $|\delta|$ and $n_1=n_2$. The overall impression created by the side-by-side display of the numerically computed variances and the Monte Carlo mean of the sample-estimated variances in Table III is that the mean sample-estimated variance approximates the magnitude of the true variance reasonably well when $|\delta| > 0$. However, sample estimates of variance based on the method of statistical differentials generally underestimate the true sampling variability, and this pattern is clearly apparent when we examine the standardized bias and skewness of the sampling distribution of this point estimator for the variance of $\hat{OVL}$.

The standardized bias in Table III is simply the discrepancy between the mean sample-estimated variance and the numerically-computed variance of $\hat{OVL}$ in terms of the Monte Carlo standard deviation of the sample-estimated variances. The usual one-sample test statistic for equality of the true mean sample-estimated variance and the numerically computed variance can be obtained from the standardized bias; the observed discrepancy is statistically different at the 0.00001 level of significance for all combinations of $|\delta|$ and $n_1=n_2$ except when $n_1=n_2 = 100$ and $n_1=n_2 = 200$ for $|\delta| = 2.00$ and $|\delta| = 3.00$. More interesting, however, is the difference between the pattern of bias exhibited by the approximation (9) as point estimator in Table III and the pattern relative error of this approximation in Table I when the true parameter values are known.

TABLE III

THE VARIANCE APPROXIMATION AS POINT ESTIMATOR FOR VAR($\hat{OVL}$)

$n_1 = n_2$	Computed Variance	Sample Estimated Variance: Mean	Sample Estimated Variance: Standard Deviation	Standardized Bias	Percent Underestimated
		$\|\delta\| = 0.00$, OVL=1.000000			
25	0.0049450734	0.0072221102	0.0022901175	0.994288	19.94
50	0.0023912376	0.0035890939	0.0011511519	1.040572	14.60
100	0.0011759309	0.0018059886	0.0005808055	1.084800	10.14
200	0.0005831244	0.0009001033	0.0002921977	1.084810	6.66
		$\|\delta\| = 0.25$, OVL=0.900524			
25	0.0080095090	0.0080616673	0.0024917009	0.020933	50.26
50	0.0048161195	0.0044570553	0.0013113821	−0.273806	52.06
100	0.0028555658	0.0025485127	0.0006112883	−0.502305	52.26
200	0.0015592059	0.0014420960	0.0002144109	−0.546194	64.26
		$\|\delta\| = 0.50$, OVL=0.802587			
25	0.0116542414	0.0098374928	0.0021629982	−0.839921	89.20
50	0.0062159275	0.0056357571	0.0007229526	−0.802501	100.00
100	0.0031236787	0.0030230848	0.0001470609	−0.684028	100.00
200	0.0015523322	0.0015383211	0.0000213639	−0.655834	72.40
		$\|\delta\| = 0.75$, OVL=0.707660			
25	0.0122534972	0.0108058992	0.0012722867	−1.137792	100.00
50	0.0060578696	0.0057953034	0.0002882812	−0.910799	93.62
100	0.0029947719	0.0029435816	0.0000882300	−0.580192	68.96
200	0.0014886710	0.0014763607	0.0000314439	−0.391500	62.66
		$\|\delta\|$=1.00, OVL=0.617075			
25	0.0115451646	0.0106714860	0.0009715522	−0.899261	81.32
50	0.0056783756	0.0054970559	0.0003511855	−0.516307	66.04
100	0.0028141844	0.0027743206	0.0001179860	−0.337869	61.42
200	0.0014007817	0.0013898151	0.0000425984	−0.257442	58.20
		$\|\delta\|$=1.50, OVL=0.453255			
25	0.0093225263	0.0088848812	0.0014181568	−0.308601	57.98
50	0.0046590311	0.0045454982	0.0004991198	−0.227466	56.48
100	0.0023272049	0.0022986755	0.0001738639	−0.164091	53.82
200	0.0011628133	0.0011562967	0.0000605768	−0.107576	52.36
		$\|\delta\|$=2.00, OVL=0.317311			
25	0.0067567611	0.0066103051	0.0016143300	−0.090722	52.58
50	0.0034499564	0.0033904909	0.0005635923	−0.105512	53.06
100	0.0017412907	0.0017339412	0.0002024666	−0.036300	51.44
200	0.0008745164	0.0008719470	0.0000722214	−0.035577	50.02
		$\|\delta\|$=3.00, OVL=0.133614			
25	0.0024961845	0.0026573542	0.0012561824	0.128301	48.82
50	0.0013354025	0.0013770637	0.0004555266	0.091457	48.92
100	0.0006901388	0.0007002007	0.0001638920	0.061394	49.46
200	0.0003507254	0.0003539061	0.0000590043	0.053905	49.10

Note: The mean and standard deviation in each row are obtained from 5,000 Monte Carlo realizations of the sample estimated variance using the approximation formula (9).

The obvious departure occurs where the standard difference $|\delta|$ is zero; the mean sample-estimated variance is approximately one Monte Carlo standard deviation greater than the computed variance of $\hat{OVL}$, precisely where the variance approximation itself demonstrates its greatest negative relative error. As $|\delta|$ increases, the bias of the sample-estimated variance declines to its maximum negative bias between $|\delta| = 0.50$ and $|\delta| = 1.00$. The magnitude of the negative bias then begins to decrease gradually until the bias once again becomes positive (or zero for large sample sizes) at $|\delta| = 3.00$. The explanation for this behavior can be traced to the fact that the sample-estimated variance in a nonlinear function (itself an approximation to the true variance) of a biased estimator for the true standard distance $|\delta|$ based on $|\bar{X}_1 - \bar{X}_2|$.

In addition to the evident bias of the sample-estimated variance, the Monte Carlo simulation indicates the skewed sampling distribution of this point estimator for the true variance of $\hat{OVL}$. To illustrate this behavior, we include in Table III the percentage of Monte Carlo realizations of the sample-estimated variance that underestimate the numerically-computed variance. The skewness of the sampling distribution is directly related to the bias of the sample-estimated variance, and the skewness toward small values of the sample-estimated variance becomes extreme at the point of greatest negative bias, where, regardless of the sample sizes, all realizations of the sample-estimated variance of $\hat{OVL}$ underestimate the numerically-computed variance. Thus the variance approximation (9) when used with sample estimates for μ_1, μ_2 and σ^2 proves valuable more as an indicator of the magnitude of the true sampling variation than as a reliable point estimate for the true variance of $\hat{OVL}$.

Alternative point estimators for OVL

The maximum-likelihood point estimator $\hat{OVL}$ given in (3) can be modified to produce other point estimators for the true overlap between two normal densities with equal variances. Two such estimators can be written in the general form

$$\tilde{OVL} = 2\,\Phi\left(-\frac{C^{\frac{1}{2}}\,|\bar{X}_1 - \bar{X}_2|}{2S}\right). \tag{10}$$

When we set $C = \left(\frac{n_1+n_2-2}{n_1+n_2}\right)$, we in effect use the unbiased point estimator for the common variance σ^2 instead of the maximum-likelihood estimator. We observe that the modified point estimator for OVL retains the downward bias of $\hat{OVL}$, since the bias of $|\bar{X}_1 - \bar{X}_2|$ dominates the behavior of $\tilde{OVL}$ as well as $\hat{OVL}$. When sample sizes are large the difference between these point estimators becomes trivial.

We have noted the relationship between OVL and the misclassification probability in the two population classification problem. Setting $C = \left(\frac{n_1+n_2-4}{n_1+n_2-2}\right)$ in the expression for the modified point estimator follows the suggestion of Anderson (1984, p. 221) that the less biased estimator for the Mahalanobis distance thus obtained provides a less biased estimator for the misclassification probability, and therefore a less biased estimator for OVL. The distribution function of the point estimators given by (10) can obviously be formulated in terms of the noncentral F distribution function, like that of $\hat{OVL}$ in (4).

More elaborate point estimators can be constructed by adding terms to "correct" for the bias of $\hat{OVL}$ or $\tilde{OVL}$. For example, we can use the implicit bias in the first-order Taylor-series representation of $\hat{OVL}$ to modify $\hat{OVL}$. A more complicated estimator can be based on the quadratic representation of the estimated misclassification probability in terms of the sample Mahalanobis distance due to McLachlan, also cited in Anderson (1984).

The Unequal Variance Case

When the two normal distributions have unequal variances, the two densities $f_1(X;\mu_1, \sigma_1^2)$ and $f_2(X;\mu_2, \sigma_2^2)$ cross at two points; see Figure 3. The values of X at these points of intersection can be determined analytically as the following:

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

Let X_1 denote the smaller of these numerical values and X_2 denote the larger. Then OVL can be computed from equation 1 as

$$OVL = \Phi\left(\frac{X_1-\mu_1}{\sigma_1}\right) + \Phi\left(\frac{X_2-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{X_1-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{X_2-\mu_1}{\sigma_1}\right) + 1 .$$

For the values of $\mu_1=0$, $\mu_2=1$, $\sigma_1^2=1$, and $\sigma_2^2=4$, shown in Figure 3, $X_1 = -1.84754$, $X_2=1.18088$, and $OVL=0.60993$. We note that the limiting value for OVL when $\sigma_1^2\neq\sigma_2^2$ as $\sigma_2^2\rightarrow\sigma_1^2$ is equation 2, the value of OVL when the two variances are equal. As in the equal variance situation, the maximum-likelihood estimator for OVL when $\sigma_1^2\neq\sigma_2^2$ can be obtained by substituting the appropriate maximum-likelihood estimators for the

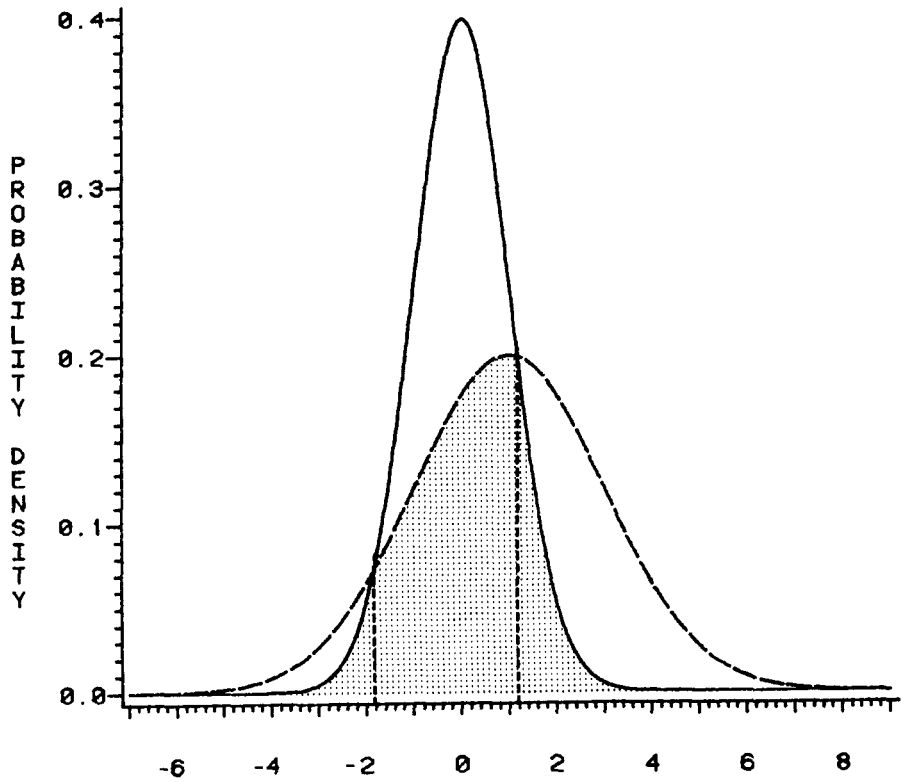

FIG. 3. The overlap between two normal distributions with unequal variances. The two points of intersection are indicated by the vertical broken lines.

parameters μ_1, μ_2 , σ_1^2 , and σ_2^2 into the expression for OVL. An approximation to the variance of this estimator has been derived (Inman, 1984, equation 2.43), and the sampling properties of $\hat{\text{OVL}}$ when $\sigma_1^2 \neq \sigma_2^2$ are similar to those described above for $\hat{\text{OVL}}$ when $\sigma_1^2 = \sigma_2^2$, particularly in respect to the downward bias of $\hat{\text{OVL}}$. For more detail, the reader should consult Inman (1984).

AN EXAMPLE OF THE USE OF THE OVERLAPPING COEFFICIENT

Here we present an example of the use of the overlapping coefficient in the context of interest to us, where tests of statistical significance are not particularly meaningful. We will ignore the results obtained above and suppose that the normality of

TABLE IV

COMPARISIONS OF THE SAMPLING DISTRIBUTION OF $\hat{OVL}$ TO THE NORMAL DISTRIBUTION WITH MONTE CARLO MEAN AND VARIANCE

$n_1=n_2$	Overlapping Coefficient	Kolmogorov Smirnov	$n_1=n_2$	Overlapping Coefficient	Kolmogorov Smirnov
	$\|\delta\|$ = 0.00, OVL=1.00000			$\|\delta\|$= 1.00, OVL=0.61708	
25	0.7993676	0.091858*	25	0.9699051	0.005316
50	0.7999142	0.090333*	50	0.9814399	0.007179
100	0.8014295	0.093619*	100	0.9839986	0.005397
200	0.8005341	0.092469*	200	0.9894579	0.004985
	$\|\delta\|$ = 0.25, OVL=0.90052			$\|\delta\|$= 1.50, OVL=0.45325	
25	0.8363752	0.078582*	25	0.9530775	0.012781*
50	0.8813634	0.057854*	50	0.9703033	0.006178
100	0.9433063	0.027406*	100	0.9788199	0.006118
200	0.9880724	0.008418*	200	0.9840617	0.005170
	$\|\delta\|$ = 0.50, OVL=0.80259			$\|\delta\|$= 2.00, OVL=0.31731	
25	0.9492364	0.027832*	25	0.9363575	0.015994*
50	0.9824978	0.005659	50	0.9562901	0.011891*
100	0.9865127	0.005921	100	0.9714937	0.007842*
200	0.9919372	0.003654	200	0.9740401	0.006953
	$\|\delta\|$ = 0.75, OVL=0.70766			$\|\delta\|$= 3.00, OVL=0.13361	
25	0.9743233	0.005073	25	0.8977146	0.039246*
50	0.9851645	0.004942	50	0.9215373	0.029631*
100	0.9896637	0.004636	100	0.9467540	0.022852*
200	0.9891637	0.004576	200	0.9590682	0.013673*

Note: All overlapping coefficients are computed using finite-difference approximations to the derivative of the theoretical distribution function (5). Each Kolmogorov-Smirnov test is based on the sample distribution function obtained from 20,000 Monte-Carlo ' realizations. Asterisks indicate rejection at the 0.01 (or less) level of significance using the pseudocritical value in Stephans (1974) for testing normality when mean and variance are unknown.

the sampling distribution of $\hat{OVL}$ based on independent samples from two normal densities with equal variances is to be investigated in a simple Monte Carlo simulation. For the values of $|\delta|$ and $n_1=n_2$ in Table IV, we have generated 20,000 realizations of $\hat{OVL}$. From these realizations we calculate the Monte Carlo mean and variance and the Kolmogorov-Smirnov test-statistic for normality based on the distribution function obtained from the 20,000 realizations of $\hat{OVL}$ and the Monte Carlo mean and variance. The value of the Kolmogorov-Smirnov test-statistic computed for each value of $|\delta|$ and $n_1=n_2$ can be found in Table IV. In addition, we again calculate the overlapping coefficient between the theoretical density of $\hat{OVL}$ (obtained by the finite-difference approximation to the derivative of equation 5) and the normal density, here specified by the Monte Carlo mean and variance. In Table IV, based on the Kolmogorov-Smirnov

test and the pseudocritical value given by Stephans (1974) for a test at the 0.01 level of significance, we would reject the hypothesis that the sampling distribution of $\hat{OVL}$ is normal for $n_1=n_2 = 200$ and $|\delta| = 0.25$. However, the overlap of the two densities here exceeds the overlap observed for eleven other combinations of $n_1=n_2$ and $|\delta|$ where we would not reject the hypothesis of normality on the basis of the Monte Carlo distribution function. Our point is not that the statistical test for normality is wrong. Instead this admittedly extreme example simply shows that conclusions, in this case that two distributions differ, reached only through the use of statistical tests can hide the underlying similarity of the phenomena of interest. These calculations serve as a reminder of our motivation for proposing the overlapping coefficient as a measure of agreement in large-sample problems where the criterion of statistical significance can become misleading and as an example of computing the overlapping coefficient in the more interesting general case.

DISCUSSION

The overlapping coefficient possesses three notable advantages as a measure of the agreement between two distributions. First, it provides a common approach for the measure of the similarity of these distributions in any distributional setting. Second, OVL is based on a simple, easily comprehended concept of the agreement or similarity of probability distributions. Third, the invariance of OVL under appropriate transformation makes this measure of agreement attractive from the standpoint of computation and estimation. One obvious weakness, as Gastwirth (1975) notes, is that the magnitude of OVL in itself does not indicate where the common probability mass is located; thus comparisons of several pairs of distributions related quite differently can still yield equal numerical values of OVL.

The sampling behavior of the maximum-likelihood estimator for the overlap between two normal densities with equal variances demonstrates that point estimators for OVL can provide useful, reliable insight and exhibit familiar statistical properties when sample sizes are large. However, point estimation may prove hazardous when sample sizes are small, especially when the two distributions of interest have very similar densities. Because in this distributional setting OVL and $\hat{OVL}$ are monotonic transformations of the Mahalanobis distance, statisticians may regard the overlapping coefficient as redundant. Nevertheless, $\hat{OVL}$ can still prove useful here or in other situations when the interpretation, as well as the computation, of $\hat{OVL}$ is conditioned on the observed sample

realizations, treating $\hat{OVL}$ as OVL computed from distributions with parameter values equal to their sample estimates. Instead of using $\hat{OVL}$ to test the equivalence of two distributions, we believe $\hat{OVL}$ supplements ordinary statistical tests for the equality of two distributions. That is, $\hat{OVL}$ can temper the conclusions drawn from such tests, particularly when the test procedures possess high statistical power and the distributions of interest are not identical, but are in fact not very different. Thus the overlapping coefficient serves as one method of assessing practical meaning or significance of differences between distributions declared "statistically significant" by another technique.

Confidence interval estimates for OVL (and D, of course) in the case of normal populations can be developed from the results we have presented here. One approach, unfortunately flawed as presented in Mishra et al. (1986), is based on the sampling properties of the sample standard distance in the expression for $\hat{OVL}$ (3). Another obvious method is based on the asymptotic normality of $\hat{OVL}$ and the estimated standard error, using the approximate variance formula (9). Procedures for testing various hypotheses involving OVL can be related to well known statistical tests involving the parameters $\mu_1 - \mu_2$, δ, or δ^2.

BIBLIOGRAPHY

Abramowitz, M. and Stegun, I.A., ed. (1972). Handbook of Mathematical Functions, Department of Commerce, National Bureau of Standards, Applied Mathematics Series, No. 55, 10th printing with corrections. Washington: Government Printing Office, 949, eq. 26.7.10.

Ahmad, I.A. (1985). "Matusita Distance," Encyclopedia of Statistical Sciences, 5. New York: John Wiley, 334-336.

Anderson, T.W. (1984). An Introduction to Multivariate Statistical Analysis. 2d ed. New York: John Wiley, 204-223.

Boring, E.G. (1919). "Mathematical vs. Scientific Significance," Psychological Bulletin, 16, 335-338.

Box, G.E.P. and Cox, D.R. (1964). "An Analysis of Transformations," Journal of the Royal Statistical Society, ser. B, 26, 211-252.

Bradley, E.L. (1985). "Overlapping Coefficient," Encyclopedia of Statistical Sciences, 6. New York: John Wiley, 546-547.

Bradley, E.L. and Piantadosi, S. (1982). The Overlapping Coefficient as a Measure of Agreement Between Distributions. Technical Report. Department of Biostatistics and Biomathematics, University of Alabama at Birmingham.

Cortese, C.F., Falk, R.F. and Cohen J.K. (1976). "Further Considerations on the Methodological Analysis of Segregation Indices," American Sociological Review, 41, 630-637.

Duncan, O.D. and Duncan B. (1955). "A Methodological Analysis of Segregation Indices, American Sociological Review, 20, 210-217.

Elandt, R.C. (1961). "The Folded Normal Distributions: Two Methods of Estimating Parameters from Moments," Technometrics, 3, 551-562.

Flury, B.K. and Riedwyl, H. (1986). "Standard Distance in Univariate and Multivariate Analysis," American Statistician, 40, 249-251.

Gastwirth, J.L. (1975). "Statistical Measures of Earning Differentials," American Statistician, 29, 32-35.

Gower, J.C. (1985). "Measures of Similarity, Dissimilarity, and Distance," Encyclopedia of Statistical Sciences , 5. New York: John Wiley, 402-403.

Ibragimov, I.A. and Has'minskii, R.Z. (1981). Statistical Estimation (Asymptotic Theory). Applications of Mathematics, Vol. 6. New York: Springer-Verlag.

Inman, H.F. (1984). Behavior and Properties of the Overlapping Coefficient as a Measure of Agreement between Distributions. Doctoral dissertation, University of Alabama in Birmingham.

Johnson, N.L. and Kotz, S. (1970). Continuous Univariate Distributions, 2. New York: John Wiley, 136-137, 170, 197.

Kamps, U. (1988). "Distance Measures in a One-parameter Class of Density Functions," Communications in Statistics -- Theory and Methods, 17, 2013-2019.

Kendall, M. and Stuart, A. (1977). The Advanced Theory of Statistics, I: Distributions Theory. 4th ed. New York: Macmillan, 246-247.

Kullback, S. (1983). "Kullback Information," Encyclopedia of Statistical Sciences, 4. New York: John Wiley, 421-425.

Leone, F.C., Nelson, L.S. and Nottingham, R.B. (1961). "The Folded Normal Distribution," Technometrics, 3, 543-550.

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

Marx, W. (1976b). "Die statistische Sicherung des Überlappungs-Koeffizienten," Zeitschrift für experimentelle und angewandte Psychologie, 23, 267-70.

Mishra, S.N., Shah, A.K. and Lefante, J.J. (1986). "Overlapping Coefficient: The Generalized t Approach," Communications in Statistics -- Theory and Methods, 15, 123-128.

Pearson, K. (1895). "Contributions to the Mathematical Theory of Evolution, II: Skew Variation in Homogeneous Material," Philosophical Transactions of the Royal Society of London, ser. A, 186, 343-414.

Reiser, B. and Guttman, I. (1986). "Statistical Inference for Pr(Y>X): The Normal Case," Technometrics, 28, 253-257.

Sneath, P.H.A. (1977). "A Method for Testing the Distinctness of Clusters: A Test of the Disjunction of Two Clusters in Euclidean Space as Measured by Their Overlap," Mathematical Geology, 9, 123-143.

Sneath, P.H.A. (1979). "The Sampling Distribution of the W Statistic of Disjunction for the Arbitrary Division of a Random Rectangular Distribution." Mathematical Geology, 11, 423-442.

Stephans, M.A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons," Journal of the American Statistical Association, 69, 730-737.

Tukey, J.W. (1957). "On the Comparative Anatomy of Transformations", Annals of Mathematical Statistics, 28, 602-632.

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.

Winship, C. (1977). "Reevaluation of Indexes of Residential Segregation," Social Forces, 55, 1058-1066.

Received March 1988; Revised July 1989.

Recommended by Irwin Guttman, University of Toronoto, CANADA.

Refereed Anonymously.