AN EVALUATION OF SMOOTHED ERROR RATE ESTIMATORS IN DISCRIMINANT ANALYSIS

by

Steven M. Snapinn

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1438

May 1983

AN EVALUATION OF SMOOTHED ERROR RATE ESTIMATORS IN DISCRIMINANT ANALYSIS

by

Steven M. Snapinn

A Dissertation submitted to the Faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill

1983

Approved by:

Adviser

Reader

Reader

STEVEN MICHAEL SNAPINN. An Evaluation of Smoothed Error Rate Estimators in Discriminant Analysis (Under the direction of JAMES D. KNOKE.)

No definitive technique for the estimation of error rates in discriminant analysis currently exists. The most commonly used estimators include the parametric plug-in estimator (D method), and the nonparametric resubstitution estimator (R method) and leave-one-out estimator (U method). The U method has received much attention recently because it is less biased than any other method; however, Glick (1978) has shown it to have very large variance. Smoothed error rate estimators, suggested by Glick in order to reduce the variance of the R method, are evaluated and compared with several other methods of estimation.

Estimators are compared under a variety of conditions, including univariate normal, multivariate normal, and some non-normal observation vectors; and several combinations of training sample sizes and prior probabilities. The basis of comparison among the estimators, unconditional mean square error, is calculated in the univariate normal situations using both numerical integration and Monte Carlo sampling, and in all other situations using only Monte Carlo sampling.

Smoothing is found to reduce the variance of the R method considerably; the greater the level of smoothing the larger the variance reduction. Low levels of smoothing tend to reduce the normally optimistic bias of the R method, whereas high levels tend to introduce a pessimistic bias. An algorithm is presented for determining a reasonable level of smoothing, as a function of the training sample sizes and the number of dimensions of the observation vector. In terms of unconditional mean square error, the particular smoothed

error rate estimator defined in this way, called the NS method, is a nonparametric estimator which is generally better than the R and U methods, and which is nearly as good as existing parametric estimators under normality.

The use of the NS method is illustrated on a data set collected to study two-year survival following myocardial infarction. A FORTRAN subprogram is presented which may be used to evaluate the NS estimate for several different types of discriminant functions.

ACKNOWLEDGEMENTS

The preparation of this dissertation involved an enormous amount of work, and I am grateful to the people who were of most help. First, I wish to thank my adviser, Dr. James Knoke, for his assistance in defining the topic, formulating a research plan, and, especially, in the actual writing of the paper. Jim has had an immeasurable impact on the organization, appearance, and professional style of this dissertation. I hope that in my later work I demonstrate much of this ability. I would also like to thank my committee, consisting of Clarence E. Davis, Gerardo Heiss, Ronald Helms, Kerry Lee and Victor Schoenbach, for the time and interest they put into this project.

Without the guidance of Jim Knoke and my committee, this dissertation would be of much lower quality than it is, but without the motivation supplied by my wife, Sylvia Wallace, it might never have been written at all. Aside from her effort in typing this paper, which was considerable, her desire to have me finish this dissertation and assume the financial responsibilities for our family found many ways to keep me on track. Most recently, the imminent prospect of an addition to our family has provided ample incentive.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. INTRODUCTION

I.1 A History of the Problem of Error Rate Estimation in Discriminant Analysis

I.1.1 Discriminant analysis

Discriminant analysis may be performed with any number of groups; however, for convenience the two group problem is considered here. Let π_1 and π_2 refer to two mutually exclusive parent populations, each having a distribution function from the same parametric class. The r-dimensional random variable, $\underset{\sim}{X}$, is defined such that the distribution function of $\underset{\sim}{X}$ is $F(\underset{\sim}{x};\underset{\sim}{\theta}_1)$ when $\underset{\sim}{X}$ is from π_1 and $F(\underset{\sim}{x};\underset{\sim}{\theta}_2)$ when $\underset{\sim}{X}$ is from π_2; where $\underset{\sim}{\theta}_i$, i=1,2, is a vector of parameters. Given an observation, $\underset{\sim}{x}$, the purpose of discriminant analysis is to classify the observation into its population of origin. Define q_i as the prior probability that an observation belongs to π_i. Similarly, let c_i be the cost of misclassifying an observation belonging to π_i, and let p_i be the probability that an observation from π_i is misclassified. The expected loss, or risk, for an observation from π_i is the product of the cost of a mistake and the probability of making it: $p_i c_i$. The expected loss due to misclassification for a random observation is the sum over the two populations of the product of the probability of drawing from the population and the expected loss for the population: $q_1c_1p_1 + q_2c_2p_2$ (Anderson 1951).

When $\underset{\sim}{\theta} = \left(\underset{\sim}{\theta}_1 \vdots \underset{\sim}{\theta}_2 \right)$ is known, the Bayes rule which minimizes the loss due to misclassification is based on the likelihood ratio (Wald 1944). Let the likelihood ratio be

$$\ell_{\underset{\sim}{\theta}}(\underset{\sim}{x}) = f(\underset{\sim}{x};\underset{\sim}{\theta}_1) \;/\; f(\underset{\sim}{x};\underset{\sim}{\theta}_2)$$

and classify $\underset{\sim}{x}$ as follows:

$$\text{if } \ell_{\underset{\sim}{\theta}}(\underset{\sim}{x}) \begin{cases} > k \text{ classify } \underset{\sim}{x} \text{ into } \pi_1 \\ < k \text{ classify } \underset{\sim}{x} \text{ into } \pi_2 \\ = k \text{ classify } \underset{\sim}{x} \text{ into } \pi_1 \text{ with probability } \gamma \\ \qquad \text{and into } \pi_2 \text{ with probability } 1-\gamma, \end{cases} \tag{1.1}$$

where $k = q_2c_1 \;/\; q_1c_2$. The symbol ζ is used to denote the rule defined by (1.1). Throughout this paper $\gamma = 0$ is assumed, and in this situation, the rule, ζ, is nonrandomized.

In most applications of discriminant analysis, $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\theta}_2$ are not known, but rather must be estimated. Samples of size n_1 and n_2, called the training samples, are chosen from π_1 and π_2, and estimates $\hat{\underset{\sim}{\Theta}}_1 = \hat{\underset{\sim}{\theta}}_1$ and $\hat{\underset{\sim}{\Theta}}_2 = \hat{\underset{\sim}{\theta}}_2$ are obtained. Note that $\hat{\underset{\sim}{\Theta}}_1$ and $\hat{\underset{\sim}{\Theta}}_2$ refer to random variables, whereas $\hat{\underset{\sim}{\theta}}_1$ and $\hat{\underset{\sim}{\theta}}_2$ refer to realizations of these random variables.

The usual procedure is to replace $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\theta}_2$ by $\hat{\underset{\sim}{\theta}}_1$ and $\hat{\underset{\sim}{\theta}}_2$ in the likelihood ratio; hence

$$\ell_{\hat{\underset{\sim}{\theta}}}(\underset{\sim}{x}) = f(\underset{\sim}{x};\hat{\underset{\sim}{\theta}}_1) \;/\; f(\underset{\sim}{x};\hat{\underset{\sim}{\theta}}_2),$$

where $\hat{\underset{\sim}{\theta}} = \left(\hat{\underset{\sim}{\theta}}_1 \vdots \hat{\underset{\sim}{\theta}}_2 \right)$. Now classify $\underset{\sim}{x}$ as follows:

$$\text{if } \ell_{\hat{\underset{\sim}{\theta}}}(\underset{\sim}{x}) \begin{cases} > k & \text{classify } \underset{\sim}{x} \text{ into } \pi_1 \\ \leq k & \text{classify } \underset{\sim}{x} \text{ into } \pi_2, \end{cases} \tag{1.2}$$

and use the symbol $\hat{\zeta}$ to denote the rule defined by (1.2). Although this rule does not minimize the loss due to misclassification, it is simple to apply and "it seems to be a reasonable procedure" (Anderson 1951).

I.1.2 Error rates of the discriminant rules

There are several different error rates associated with the discriminant rules above, and it is necessary to be very careful to avoid confusing them. First, let $\alpha_i(\zeta)$, i=1,2, be defined as the probability that a random member of π_i is misallocated when the rule ζ is used:

$$\alpha_1(\zeta) = \Pr\{\ell_{\underset{\sim}{\theta}}(\underset{\sim}{X}) \leq k \mid \underset{\sim}{X} \in \pi_1\}$$
$$\alpha_2(\zeta) = \Pr\{\ell_{\underset{\sim}{\theta}}(\underset{\sim}{X}) > k \mid \underset{\sim}{X} \in \pi_2\}.$$

These are known as the optimum error rates; they are the error rates that would occur if $\underset{\sim}{\theta}$ were known.

Since π_1 and π_2 are labelled arbitrarily, it is necessary only to consider $\alpha_1(\zeta)$. To study $\alpha_2(\zeta)$, the labels of the populations are simply interchanged. Therefore, subsequently, any unknown observation, $\underset{\sim}{X}$, is assumed to come from π_1, the subscript on α is dropped, and $\alpha(\zeta) = \alpha_1(\zeta)$. The optimum error rate is now given by

$$\alpha(\zeta) = \Pr\{\ell_{\underset{\sim}{\theta}}(\underset{\sim}{X}) \leq k\}.$$

The conditional actual error rate is defined as the probability that a random observation from π_1 is misallocated when the rule $\hat{\zeta}$ is used:

$$\alpha(\hat{\zeta}) = \Pr \{\ell_{\hat{\theta}}(\underset{\sim}{X}) \leq k\}.$$

Note that this error rate is conditional on the estimated parameters, which in turn are determined by the training samples.

The final error rate considered here is the expected actual error rate. This is the probability that randomly chosen training samples yield a decision rule which misclassifies a randomly chosen member of π_1. If the expected value operator is defined with respect to all possible training samples, then the expected actual error rate is written as

$$E[\alpha(\hat{\zeta})] = E[\Pr \{\ell_{\hat{\theta}}(\underset{\sim}{X}) \leq k\}] = \Pr \{\ell_{\hat{\Theta}}(\underset{\sim}{X}) \leq k\}.$$

Note the hierarchy associated with these error rates: the optimum error rate is a function only of the distributions of $\underset{\sim}{X}$ for the two populations, the expected actual error rate is a function of the distributions of $\underset{\sim}{X}$ and the training sample sizes, while the conditional actual error rate is a function of the distributions of $\underset{\sim}{X}$ and the particular training samples selected.

In order to compare error rate estimators it is necessary to specify the error rate being estimated. Assuming $\underset{\sim}{\theta}$ is unknown, estimates of the optimum error rate and the expected actual error rate are valuable for deciding whether or not a discriminant analysis should be performed, for comparing possible discriminant rules, and for determining the advantages of increasing the size of the training samples. However, an experimenter is most likely to be concerned with the performance of his or her discriminant rule after the training samples have been selected. Although the performance of the rule can vary greatly with the choice of the training samples, the optimum error rate and the expected actual error rate are independent of that choice. Therefore, once a discriminant rule, $\hat{\zeta}$, has been determined, it is the conditional actual error rate, $\alpha(\hat{\zeta})$, which is of interest.

I.1.3 Expressions for $\alpha(\zeta)$, $\alpha(\hat{\zeta})$ and $E[\alpha(\hat{\zeta})]$ under normality

Throughout this paper the costs of misclassification are assumed to be equal; this may be done without loss of generality since this assumption does not restrict the range of the constant k (1.1). Now consider the situation where π_1 and π_2 refer to r-variate normal parent distributions with unknown means, $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$, respectively, a common covariance matrix, $\underset{\sim}{\Sigma}$, which may be known or unknown, and let

$$\Delta^2 = (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2) \tag{1.3}$$

be the Mahalanobis distance between the populations. Also assume equal prior probabilities and, therefore, $k = 1$.

Now let $\bar{\underset{\sim}{X}}_1 = \bar{\underset{\sim}{x}}_1$, $\bar{\underset{\sim}{X}}_2 = \bar{\underset{\sim}{x}}_2$, and $\hat{\underset{\sim}{\Sigma}} = \underset{\sim}{S}$ be the minimum variance unbiased estimates of $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$, and $\underset{\sim}{\Sigma}$ based on the training samples (Anderson 1951). Note that $\hat{\underset{\sim}{\Sigma}}$ refers to a random variable, and $\underset{\sim}{S}$ to a realization of that random variable. In this situation, the linear discriminant function, or Anderson's W statistic, is defined as

$$W(\underset{\sim}{X}) = \begin{cases} [\underset{\sim}{X}-\tfrac{1}{2}(\bar{\underset{\sim}{X}}_1+\bar{\underset{\sim}{X}}_2)]'\,\underset{\sim}{\Sigma}^{-1}\,(\bar{\underset{\sim}{X}}_1-\bar{\underset{\sim}{X}}_2) & \text{if } \underset{\sim}{\Sigma} \text{ known} \\ [\underset{\sim}{X}-\tfrac{1}{2}(\bar{\underset{\sim}{X}}_1+\bar{\underset{\sim}{X}}_2)]'\,\hat{\underset{\sim}{\Sigma}}^{-1}\,(\bar{\underset{\sim}{X}}_1-\bar{\underset{\sim}{X}}_2) & \text{if } \underset{\sim}{\Sigma} \text{ unknown,} \end{cases} \tag{1.4}$$

and the decision rule, $\hat{\zeta}$ of (1.2), reduces to

$$\text{if } W(\underset{\sim}{x}) \begin{cases} > 0 & \text{assign } \underset{\sim}{x} \text{ to } \pi_1 \\ \leq 0 & \text{assign } \underset{\sim}{x} \text{ to } \pi_2. \end{cases}$$

The optimum error rate is simply

$$\alpha(\zeta) = \Phi(-\Delta/2) \tag{1.5}$$

where $\Phi(t) = \int_{-\infty}^{t} (2\pi)^{-\frac{1}{2}} e^{-x^2/2}\, dx$ (John 1961).

Conditional on the training samples (and therefore on $\bar{\underset{\sim}{x}}_1$, $\bar{\underset{\sim}{x}}_2$, and $\underset{\sim}{S}$), $W(\underset{\sim}{X})$ has a univariate normal distribution:

$$W(\underset{\sim}{X}) \sim \begin{cases} N\{ [\underset{\sim}{\mu}_1 - \tfrac{1}{2}(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2)]' \underset{\sim}{\Sigma}^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2), \\ \qquad (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' \underset{\sim}{\Sigma}^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) \} & \text{if } \underset{\sim}{\Sigma} \text{ known} \\ N\{ [\underset{\sim}{\mu}_1 - \tfrac{1}{2}(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2)]' \underset{\sim}{S}^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2), \\ \qquad (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' \underset{\sim}{S}^{-1} \underset{\sim}{\Sigma} \underset{\sim}{S}^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) \} & \text{if } \underset{\sim}{\Sigma} \text{ unknown.} \end{cases}$$

The conditional actual error rate is the probability that $W(\underset{\sim}{X})$ is less than or equal to zero and hence can be given as

$$\alpha(\hat{\zeta}) = \begin{cases} \Phi\left[- \dfrac{[\underset{\sim}{\mu}_1 - \frac{1}{2}(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2)]' \underset{\sim}{\Sigma}^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)}{[(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' \underset{\sim}{\Sigma}^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)]^{\frac{1}{2}}} \right] & \text{if } \underset{\sim}{\Sigma} \text{ known} \\ \Phi\left[- \dfrac{[\underset{\sim}{\mu}_1 - \frac{1}{2}(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2)]' \underset{\sim}{S}^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)}{[(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' \underset{\sim}{S}^{-1} \underset{\sim}{\Sigma} \underset{\sim}{S}^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)]^{\frac{1}{2}}} \right] & \text{if } \underset{\sim}{\Sigma} \text{ unknown} \end{cases} \tag{1.6}$$

(McLachlan 1972).

The expected actual error rate is more complicated. For $\underset{\sim}{\Sigma}$ unknown, an asymptotic distribution of $E[\alpha(\hat{\zeta})]$ was given by McLachlan (1972, 1974a), and Sayre (1980) used numerical integration to tabulate values of $E[\alpha(\hat{\zeta})]$ for $r = 1,\ldots,4$ and $n_1 = n_2 = 25$, 50, 100. These results were compared and were found to be in close agreement.

In the univariate case, $r = 1$, the situation simplifies considerably. Equation (1.5) involves only $\Delta^2 = (\mu_1-\mu_2)^2 / \sigma^2$. Equation (1.6) reduces to

$$\alpha(\hat{\zeta}) = \begin{cases} \Phi\{ -[\mu_1-\tfrac{1}{2}(\bar{x}_1+\bar{x}_2)] / \sigma \} & \text{if } \bar{x}_1 > \bar{x}_2 \\ 1 - \Phi\{ -[\mu_1-\tfrac{1}{2}(\bar{x}_1+\bar{x}_2)] / \sigma \} & \text{if } \bar{x}_1 \leq \bar{x}_2, \end{cases}$$

for σ^2 known or unknown. John (1961) showed that

$$E[\alpha(\hat{\zeta})] = G(-a,b;\rho) + G(a,-b;\rho) \qquad (1.7)$$

where
$$a = (\mu_2-\mu_1)(n_1n_2)^{\frac{1}{2}} / \sigma(n_1+n_2)^{\frac{1}{2}}$$
$$b = (\mu_2-\mu_1)(n_1n_2)^{\frac{1}{2}} / \sigma(4n_1n_2+n_1+n_2)^{\frac{1}{2}}$$
$$\rho = (n_2-n_1) / [(n_1+n_2)(4n_1n_2+n_1+n_2)]^{\frac{1}{2}}$$
and
$$G(t_1,t_2;\rho) = [2\pi(1-\rho^2)^{\frac{1}{2}}]^{-1} \int_{-\infty}^{t_1} \int_{-\infty}^{t_2} \exp\{ - x_1^2-2\rho x_1x_2+x_2^2 \;/\; 2(1-\rho^2) \}\, dx_2dx_1.$$

I.1.4 Criteria for comparing error rate estimators

Let $\hat{\alpha}$ represent an arbitrary estimator of the conditional actual error rate, $\alpha(\hat{\zeta})$, based on the training samples. The most reasonable criterion for comparing estimators is felt to be

$$E[\hat{\alpha} - \alpha(\hat{\zeta})]^2, \qquad (1.8)$$

called the unconditional mean square error (UMSE) by Sedransk and Okamoto (1971). Two other possible criteria are the conditional mean square error,

$$E\ [\{\hat{\alpha} - \alpha(\hat{\zeta})\}^2 \mid \hat{\underset{\sim}{\theta}}] , \qquad (1.9)$$

and the mean absolute error

$$E\left| \hat{\alpha} - \alpha(\hat{\zeta}) \right|. \qquad (1.10)$$

The results obtained using the criterion of conditional mean square error are functions of $\hat{\underset{\sim}{\theta}}$; this criterion could be used if it were desirable to have the choice of the error rate estimator depend on the training samples. However, the goal of this study is to compare estimators chosen independently of the training samples. Therefore, UMSE, which is the expected value of the conditional mean square error over the distribution of $\hat{\underset{\sim}{\Theta}}$, is the preferred criterion. The mean absolute error is also felt to be a reasonable criterion, but it is not considered further because it is not as sensitive to the variability of the error as the unconditional mean square error.

I.2 Error Rate Estimators

Each of the methods of error rate estimation described in this section is given a symbol to identify it. The estimators are referred to by $\hat{\alpha}$ with that symbol as a superscript.

I.2.1 Plug-in estimator

The earliest error rate estimator, called the plug-in estimator or the D method, was proposed by Fisher (1936). Let

$$D^2 = \begin{cases} (\bar{\underset{\sim}{X}}_1-\bar{\underset{\sim}{X}}_2)' \, \underset{\sim}{\Sigma}^{-1} \, (\bar{\underset{\sim}{X}}_1-\bar{\underset{\sim}{X}}_2) \text{ if } \underset{\sim}{\Sigma} \text{ known} \\ (\bar{\underset{\sim}{X}}_1-\bar{\underset{\sim}{X}}_2)' \, \hat{\underset{\sim}{\Sigma}}^{-1} \, (\bar{\underset{\sim}{X}}_1-\bar{\underset{\sim}{X}}_2) \text{ if } \underset{\sim}{\Sigma} \text{ unknown;} \end{cases} \qquad (1.11)$$

the plug-in estimator is defined by

$$\hat{\alpha}^D = \Phi(-D/2).$$

Note that this estimator can be obtained in two ways: by substituting D for Δ in (1.5); or, conditioned on $\bar{\underset{\sim}{x}}_1$, $\bar{\underset{\sim}{x}}_2$, and $\underset{\sim}{S}$, by substituting $\bar{\underset{\sim}{x}}_1$ for $\underset{\sim}{\mu}_1$ and, if $\underset{\sim}{\Sigma}$ unknown, $\underset{\sim}{S}$ for $\underset{\sim}{\Sigma}$, in (1.6).

Many investigators have found that this estimator is optimistically biased for the conditional actual error rate in that it tends to be smaller than $\alpha(\hat{\zeta})$ (Dunn and Varady 1966; Lachenbruch and Mickey 1968; Glick 1978). The distribution of $\hat{\alpha}^D$ has been well studied. In the univariate case with σ^2 known, Hills (1966) established that

$$E(\hat{\alpha}^D) = \Phi(a_1) + \Phi(b_1) - 2G(a_1,b_1;\rho_1) \tag{1.12}$$

where $a_1 = \Delta/2(1+h_1)^{\frac{1}{2}}$, $b_1 = \Delta/2(h_1)^{\frac{1}{2}}$, $\rho_1 = [h_1/(1+h_1)]^{\frac{1}{2}}$ and $h_1 = (n_1+n_2)/4n_1n_2$. Moran (1975) generalized (1.12) to the r-variate normal case with $\underset{\sim}{\Sigma}$ known:

$$E(\hat{\alpha}^D) = \Pr\,\{w_1/w_2 > (1+\rho_1)/(1-\rho_1)\}$$

where ρ_1 is as for (1.12) and w_1 and w_2 are independently distributed non-central χ^2 random variables, each with r degrees of freedom. The non-centrality parameters are complicated functions of Δ^2 and the training sample sizes. The asymptotic distribution of $\hat{\alpha}^D$ has been studied by McLachlan (1973) and Schervish (1981). This estimator has been studied using Monte Carlo simulation by Dunn (1971) and Dunn and Varady (1966), who determined confidence intervals for $\alpha(\hat{\zeta})$ and $\alpha(\zeta)$ given $\hat{\alpha}^D$, the training sample sizes, r and Δ.

I.2.2 Resubstitution estimator

Another commonly used error rate estimator, called the resubstitution estimator, the apparent error rate, or the R method, was first proposed by Smith (1947). This is the proportion of the observations in the training sample from π_1 which is misclassified by the discriminant rule.

Das Gupta (1974) presented the following inequalities in the r-variate normal case:

$$E(\hat{\alpha}^R) < E(\hat{\alpha}^D) \quad \text{if } \underset{\sim}{\Sigma} \text{ is known}$$

$$E(\hat{\alpha}^R) < \alpha(\zeta) < E[\alpha(\hat{\zeta})] \quad \text{if } n_1 = n_2 \text{ and } \underset{\sim}{\Sigma} \text{ is known or unknown}$$

$$E(\hat{\alpha}^R) < E[\alpha(\hat{\zeta})] \quad \text{if } \underset{\sim}{\Sigma} \text{ is known and } r = 1.$$

Although none of the above inequalities directly compares $\hat{\alpha}^R$ and $\alpha(\hat{\zeta})$, they may serve as indirect comparisons, indicating that $\hat{\alpha}^R$ is optimistically biased for $\alpha(\hat{\zeta})$; more so than $\hat{\alpha}^D$. Sampling studies (Lachenbruch and Mickey 1968; Glick 1978) have shown that the bias of the resubstitution estimator with respect to $\alpha(\hat{\zeta})$ may be very large, especially when n_1, n_2, and Δ^2 are small.

In the univariate normal case with σ^2 known, Hills (1966) established that

$$E(\hat{\alpha}^R) = \Phi(a_2) + \Phi(b_2) - 2G(a_2, b_2; \rho_2) \tag{1.13}$$

where $a_2 = \Delta / 2(h_2)^{\frac{1}{2}}$, $b_2 = \Delta / 2(1+h_2-n_1^{-1})^{\frac{1}{2}}$, $\rho_2 = [(n_1+n_2) / (4n_1n_2+n_1-3n_2)]^{\frac{1}{2}}$, and $h_2 = (n_1+n_2) / 4n_1n_2$. In the multivariate normal case with $\underset{\sim}{\Sigma}$ known, Moran (1975) found that

$$E(\hat{\alpha}^R) = \Pr\{\ w_3 / w_4 > (1+\rho_2) / (1-\rho_2)\ \}$$

where ρ_2 is as in (1.13) and w_3 and w_4 are independently distributed non-central χ^2 random variables, each with r degrees of freedom. The non-centrality parameters are complicated functions of Δ^2 and the training sample sizes. Schervish (1981) derived asymptotic expansions for both the mean and variance of $\hat{\alpha}^R$, and McLachlan (1976), working with asymptotic expansions, found an expression for the bias of the resubstitution estimator with respect to the expected actual error rate.

The R method is the simplest example of the class of nonparametric estimators, while the D method is the simplest example of the class which assumes multivariate normality. Although under normality the R method appears to have a larger optimistic bias than the D method, one would expect the nonparametric estimators to be less sensitive to departures from normality than the parametric methods. In a Monte Carlo sampling study with uniformly distributed populations, Glick (1978) in fact found the R method to be less biased than the D method.

I.2.3 Leave-one-out estimator

Define $\underset{\sim}{X}_i$, $i=1,\ldots,n_1$, as the set of observations in the training sample from π_1, and $\hat{\zeta}_i$ as the discriminant rule obtained when $\underset{\sim}{X}_i$ is omitted. The leave-one-out estimator, or the U method, is the proportion of the $\underset{\sim}{X}_i$ misclassified by the corresponding $\hat{\zeta}_i$.

The leave-one-out estimator, also referred to as cross-validation (Efron and Gong 1983) and the jackknife estimator by the BMDP Package (Jennrich and Sampson 1982), is frequently attributed to Lachenbruch (1967) (Hills 1966; Sorum 1971), but earlier references can be found (Brailovskiy 1964; Brailovskiy and Lunts 1964). This estimator "has the flavor of the estimation methods used in statistics to reduce bias that are referred to as jackknifing procedures" (Toussaint 1974). An early reference to the jackknife, although not by that name, is Quenouille (1956).

The main advantage of this method is felt to be that it obtains an unbiased estimate of the expected actual error rate for a discrimination problem with training samples of size n_1-1 and n_2 (Lachenbruch 1967). However, this does not mean that the leave-one-out estimator has small bias with respect to the conditional actual error rate, which is the error rate of interest here.

One disadvantage of this estimator is that it requires more computation than the resubstitution estimator. However, ways have been found to reduce this problem. Lachenbruch (1967) showed how all of the discriminant rules, $\hat{\zeta}_i$, $i=1,\ldots,n$, could be obtained with only one matrix inversion, using a theorem by Bartlett (1951). Fukunaga and Kessell (1971) also considered this problem. The amount of computation required with the leave-one-out estimator is not prohibitive, however, as evidenced by this method having been implemented by the computer program BMDP7M (Jennrich and Sampson 1981).

Another disadvantage of the leave-one-out estimator is its large variance. The main consideration of most investigators when comparing estimators has been the bias, but the variance is also an important factor. Glick (1978) performed a sampling experiment in order to demonstrate the importance of the variance. In the univariate normal case, he found that the bias with respect to $E[\alpha(\hat{\zeta})]$ is very small for the leave-one-out estimator, larger for the plug-in estimator and largest for the resubstitution estimator, as expected. However, he also compared the variance of the estimators and found that the leave-one-out estimator had a much larger variance than the resubstitution estimator, which in turn had a larger variance than the plug-in estimator. Unfortunately, Glick did not consider the mean square error and, hence, left unanswered whether the resubstitution estimator ever performs better than leave-one-out.

I.2.4 Smoothed estimators

Glick (1978) defined counting-type estimators as those which can be written as

$$\sum_{i=1}^{n_1} h(\underset{\sim}{X}_i) \;/\; n_1,$$

where $0 \le h(\underset{\sim}{X}) \le 1$. Therefore, for resubstitution,

$$h(\underset{\sim}{X}_i) = \begin{cases} 1 & \text{if } \underset{\sim}{X}_i \text{ is misclassified by } \hat{\zeta} \\ 0 & \text{otherwise;} \end{cases}$$

and for leave-one-out,

$$h(\underset{\sim}{X}_i) = \begin{cases} 1 & \text{if } \underset{\sim}{X}_i \text{ is misclassified by } \hat{\zeta}_i \\ 0 & \text{otherwise.} \end{cases}$$

The large variance of the counting-type estimators is due to the fact that h is an indicator function; a very small shift in $\underset{\sim}{X}_i$ can cause a large change in $h(\underset{\sim}{X}_i)$. Glick stated that the leave-one-out estimator reduces the bias of the resubstitution estimator in "the worst possible way". He claimed that it "exacerbates rather than smooths fluctuation of the ... [resubstitution] estimator". He then argued in favor of smoothed counting estimators, defined such that $h(\underset{\sim}{X}_i)$ is a smooth function of $\underset{\sim}{X}_i$.

Glick suggested replacing the discontinuous jump function by a smooth function which reflects to some degree the amount of confidence one has when deciding whether or not $\underset{\sim}{X}_i$ is misclassified. Although this suggested a class of estimators, Glick examined one member of that class in his sampling experiment. He defined, for the univariate case and $\bar{X}_1 > \bar{X}_2$,

$$g(X_i) = \begin{cases} 1 & \text{if } [X_i-(\bar{X}_1+\bar{X}_2)/2]/\hat{\sigma} < -b \\ \tfrac{1}{2}-[X_i-(\bar{X}_1+\bar{X}_2)/2]/2b\hat{\sigma} & \text{if } [X_i-(\bar{X}_1+\bar{X}_2)/2]/\hat{\sigma} \in (-b,b) \\ 0 & \text{if } [X_i-(\bar{X}_1+\bar{X}_2)/2]/\hat{\sigma} > b \end{cases}$$

with $b = \tfrac{1}{2}$. He called this a smooth slide function, with the range of the slide arbitrarily chosen to be $(-\tfrac{1}{2},\tfrac{1}{2})$.

Glick found that the smoothed estimator did indeed have a smaller bias and variance than the resubstitution estimator. Although smoothed estimators were not examined further, Glick suggested

$$g(X_i) = \Phi\{ [\tfrac{1}{2}(\bar{X}_1+\bar{X}_2)-X_i]/b\hat{\sigma} \},$$

where b is a constant, as a more sophisticated smoothing function. He concluded that:

1) "the normal plug-in estimator ... is both optimistically biased for close normal distributions [those separated by a small Mahalanobis distance] and erratically biased for non-normal distributions. It is not robust and hence should not be used in practical problems."

2) "the leave-one-out estimator is an anti-smooth modification of counting .... Hence the leave-one-out estimator should be eschewed."

3) "fortunately, the counting method can be modified to reduce bias magnitude and standard deviation simultaneously, without sacrificing robustness."

I.2.5 The DS method

There are many estimators based on the plug-in estimator which all assume multivariate normality and contain a bias correction. When Σ is unknown, D^2 is a biased estimator of Δ^2. Lachenbruch and Mickey (1968) described a consistent estimator of Δ^2 which has less bias than D^2. This estimator of Δ^2 is

$$DS = (n_1+n_2-r-3)D^2 \;/\; (n_1+n_2-2), \tag{1.14}$$

and hence the estimator of $\alpha(\hat{\zeta})$, called the DS method, is

$$\hat{\alpha}^{DS} = \Phi\{-(DS)^{\frac{1}{2}} \;/\; 2\}.$$

I.2.6 The O and OS methods

The distribution of Anderson's W statistic (1.4) is very complicated and is not known exactly. Okamoto (1963) provided an asymptotic expansion for $\Pr\{\ W(\underset{\sim}{X}) < \frac{1}{2}\Delta^2+a\Delta\ \}$ where a is a real constant. Since $\alpha(\hat{\zeta}) = \Pr\{\ W(\underset{\sim}{X}) < 0\ \}$, one could substitute an estimate of Δ^2 into Okamoto's expansion in order to estimate $\alpha(\hat{\zeta})$. Lachenbruch and Mickey (1968) suggested two such estimators: the O method is obtained by replacing Δ^2 with D^2, and the OS method is obtained by replacing Δ^2 with DS. These estimators were explicitly obtained in the univariate case with σ^2 known by Sedransk and Okamoto (1971):

$$\hat{\alpha}^{O} = \Phi\{-|\bar{X}_2-\bar{X}_1|/2\sigma\ \} + \frac{1}{8}(n_1^{-1}+n_2^{-1})\ \phi'\{-|\bar{X}_2-\bar{X}_1|/2\sigma\ \}$$

$$\hat{\alpha}^{OS} = \Phi\left[-\frac{|\bar{X}_2-\bar{X}_1|(n_1+n_2-4)^{\frac{1}{2}}}{2\sigma(n_1+n_2-2)^{\frac{1}{2}}}\right] + \frac{1}{8}(n_1^{-1}+n_2^{-1})\ \phi'\left[-\frac{|\bar{X}_2-\bar{X}_1|(n_1+n_2-4)^{\frac{1}{2}}}{2\sigma(n_1+n_2-2)^{\frac{1}{2}}}\right]$$

where $\phi'(t) = -t(2\pi)^{-\frac{1}{2}}\exp(-t^2/2)$.

I.2.7 Other estimators

Lachenbruch (1967) and Toussaint (1974) described the holdout method. The training samples are randomly divided into two parts, not necessarily of equal size, one of which is called the holdout sample. A rule, $\hat{\zeta}^*$, is obtained from those observations not in the holdout sample. The holdout estimator is defined as the proportion of observations in the holdout sample from π_1 which is misclassified by $\hat{\zeta}^*$. The problems associated with this estimator are that fairly large training samples are required, the discriminant function which is evaluated ($\hat{\zeta}^*$) is not the same one that is used ($\hat{\zeta}$), and it is not clear how large to make the holdout sample (Lachenbruch and Mickey 1968).

Two estimators which combine aspects of the leave-one-out and holdout methods are the data-shuffling method and the rotation method. The rotation method, proposed by Toussaint (1974) is essentially a leave-n-out procedure, where the training samples have been partitioned into $(n_1+n_2)/n$ sets. In the data-shuffling method, attributed to Duda and Hart (1973), a large number of holdout sets of size n are chosen at random and with replacement.

The posterior probability estimator was described by Moore, Whitsitt and Landgrebe (1976). Assuming equal prior probabilities, if $\underset{\sim}{\theta}$ is known and the discriminant rule is ζ, the posterior probability of misclassification is

$$[\min\{ f(\underset{\sim}{X};\underset{\sim}{\theta}_1) , f(\underset{\sim}{X};\underset{\sim}{\theta}_2) \}] \;\; / \;\; [f(\underset{\sim}{X};\underset{\sim}{\theta}_1) + f(\underset{\sim}{X};\underset{\sim}{\theta}_2)] .$$

When $\underset{\sim}{\theta}$ is estimated, the posterior probability of misclassification by the rule $\hat{\zeta}$, given $\underset{\sim}{X}_i$, is estimated by

$$[\min\{ f(\underset{\sim}{X}_i;\underset{\sim}{\hat{\theta}}_1) , f(\underset{\sim}{X}_i;\underset{\sim}{\hat{\theta}}_2) \}] \ / \ [f(\underset{\sim}{X}_i;\underset{\sim}{\hat{\theta}}_1) + f(\underset{\sim}{X}_i;\underset{\sim}{\hat{\theta}}_2)] .$$

This function is evaluated for each of the $\underset{\sim}{X}_i$ and the mean is the estimator of $\alpha(\hat{\zeta})$.

Lachenbruch and Mickey (1968) defined an estimator based on the leave-one-out estimator called the $\bar{U}$ method. When $\underset{\sim}{X}_i$ is omitted and $\hat{\zeta}_i$ is calculated, D_i may also be calculated, where D_i^2 is defined as D^2 in (1.11) leaving out $\underset{\sim}{X}_i$. Let $\bar{D}$ and s_D be the mean and standard deviation of the n_1 values of D_i; the estimator based on the $\bar{U}$ method is defined as $\Phi(-\bar{D}/s_D)$.

Other estimators were defined by Broffitt and Williams (1973), Lachenbruch (1968), McLachlan (1974c), Sorum (1971 and 1973) and Toussaint (1974).

I.2.8 The bootstrap

Bootstrap methods may be applied to a variety of statistical problems; Efron (1979) showed how they may be used in discriminant analysis. This technique is mentioned here although it is actually a bias reduction method, not an estimator. The bias of the resubstitution estimator with respect to the expected actual error rate is given by

$$\delta = E(\hat{\alpha}^R) - E[\alpha(\hat{\zeta})].$$

The bootstrap method first estimates δ to arrive at an estimate of $E[\alpha(\hat{\zeta})]$: $\hat{\alpha}^R - \hat{\delta}$. Random samples of size n_1 and n_2 are generated from the distributions $F(\underset{\sim}{x};\underset{\sim}{\hat{\theta}}_1)$ and $F(\underset{\sim}{x};\underset{\sim}{\hat{\theta}}_2)$. These new samples are used to define a new rule, $\hat{\zeta}'$. Let P_{old} be the proportion of the original training sample from π_1 which is misclassified by $\hat{\zeta}'$, P_{new}

be the proportion of the new sample from π_1 which is misclassified by $\hat{\zeta}'$, and let $P = P_{new} - P_{old}$. The value of P may be used to estimate δ, but in practice the above procedure is usually repeated a large number of times and the values of P are averaged in order to reduce the variability of $\hat{\delta}$.

While the bootstrap may be used to reduce the bias of an estimator with respect to the expected actual error rate, it is the conditional actual error rate which is of interest here. The bootstrap is therefore not very valuable in this context, and is not considered further.

I.3 Comparative Studies of Error Rate Estimators

I.3.1 Monte Carlo sampling studies

Lachenbruch and Mickey (1968) performed a sampling study in which they estimated the distribution of the absolute error, $|\hat{\alpha}-\alpha(\hat{\zeta})|$, for several parametric and nonparametric estimators. Based on their results, Lachenbruch and Mickey recommended the leave-one-out or U method when the assumption of normality is questionable, but no method was found uniformly best when normality is assumed. The plug-in and resubstitution methods were found to be poor in general; of the others the OS method was found to be preferable to the O method and the DS method to the D method. In general, the OS and U methods were found to be among the best.

Sorum (1972) studied estimators of $\alpha(\zeta)$ and $E[\alpha(\hat{\zeta})]$ in the multivariate normal case with $\underset{\sim}{\Sigma}$ assumed known, using mean absolute error as the criterion. She found the DS and O methods to be among the best estimators of $\alpha(\zeta)$, followed by the U and OS methods; the R and D methods were found to be among the poorest. These groups of estimators tended to diverge as r increased. The best estimator of $E[\alpha(\hat{\zeta})]$ was found to be the OS method. The O, DS, and U methods formed a second group, followed by the R and D methods. Again, the difference between the groups increased with r. Estimators of $\alpha(\hat{\zeta})$ were also studied, reproducing the work of Lachenbruch and Mickey. The resultant ordering of estimators from best to worst was OS, O, DS, U, R and D, "closely matching" the order previously observed.

Sorum (1973) compared estimators of $\alpha(\zeta)$ and $E[\alpha(\hat{\zeta})]$ in the univariate normal case with σ^2 known, using the criterion of mean absolute error. This time the DS method stood out as the best, and the R and U methods were found to be equally poor.

Glick (1978) examined the D, R, and U methods, as well as the posterior probability estimator and his smooth slide estimator, in the univariate normal case. He compared the biases and variances of these estimators and found that the smoothed estimator did indeed have a smaller bias and variance than the resubstitution estimator. The results confirmed that the plug-in and posterior probability estimators under normality are superior to the smoothed estimator, which in turn is superior to the resubstitution estimator. It was not clear where the leave-one-out estimator fits in, since its bias was found to be much smaller, but its variance much larger, than the other estimators. Glick also repeated his sampling experiment assuming the parents to be uniformly distributed. In this case the two parametric estimators had large optimistic biases, the resubstitution and smoothed estimators had intermediate biases, and the leave-one-out estimator had a very small bias. This study demonstrated the robustness of the nonparametric estimators.

Efron (1979) performed a sampling experiment to examine how well the bootstrap method works when applied to the resubstitution estimator. With π_1 and π_2 having bivariate normal distributions with a common $\underset{\sim}{\Sigma}$, he claimed that the bootstrap outperformed the leave-one-out method. Both methods were found to have small biases, but the bootstrap method had a variance which was one-third that of the leave-one-out method.

McLachlan (1980) studied the bootstrap method and concluded that it "appears to provide an efficient estimator of the bias correction of the apparent error rate in those situations where it is most needed; that is, for underlying populations with a small Mahalanobis distance between them."

Ganesalingem and McLachlan (1980) examined the posterior probability estimator and found it to have a small optimistic bias.

I.3.2 Studies involving asymptotic expansions

Sedransk and Okamoto (1971) considered asymptotic expansions for unconditional mean square error (1.8) in the univariate normal case. They found all parametric estimators to outperform the resubstitution and leave-one-out methods. Among the parametric estimators studied, they found that when σ^2 is assumed known, the DS method is the best when $|\Delta| < 1.51$, the O method is the best when $1.51 < |\Delta| < 1.79$ and the D method is the best when $1.79 < |\Delta|$. With σ^2 assumed unknown, they found that the OS method, which they did not study with σ^2 assumed known, is the best when $|\Delta| < 1.63$; the DS method is the best when $1.63 < |\Delta| < 1.86$; the O method is the best when $1.86 < |\Delta| < 2.15$; and the D method is the best when $2.15 < |\Delta|$. They claimed that their results "are somewhat inconsistent with the results of Lachenbruch and

Mickey who showed experimentally the superiority of the estimator OS in the multivariate normal case".

Sorum (1971) considered asymptotic expansions for the conditional mean square error (1.9) in the multivariate normal case with $\underset{\sim}{\Sigma}$ assumed known. By comparing the leading terms in the expansions, she found no differences among the parametric estimators. These estimators, however, were found to be better than the R and U methods.

In the univariate normal case with σ^2 known, Sorum (1973) studied the mean square errors for estimators with respect to $\alpha(\zeta)$ and $E[\alpha(\hat{\zeta})]$. The D and DS methods were found to be the best, followed by the R method; the U method was found to be the poorest.

McLachlan (1974) examined asymptotic expansions for UMSE with $\underset{\sim}{\Sigma}$ unknown, and including a second order term, which was not done by Sorum. He did not find any estimator to be uniformly superior to the others. In general, however, for $r \geq 7$ the OS method and the M method, defined by McLachlan, were found to be the best, while for $r<7$ the OS and DS methods were found to be the best.

McLachlan (1974d) studied the relationship between asymptotic expansions for mean square errors with respect to each of $\alpha(\zeta)$, $E[\alpha(\hat{\zeta})]$ and $\alpha(\hat{\zeta})$. He explored the range of values of r and Δ for which each of the mean square errors is smallest.

McLachlan (1974c and 1975) explored the M method and directly compared it to the OS method. On the basis of asymptotic expansions of UMSE, the OS method was found to be better unless r is large and Δ^2 is small.

I.4 General Objectives of this Study

I.4.1 Evaluation of UMSE by numerical integration in the univariate normal case

The studies discussed in Section I.3 fall into one of two categories: either they used asymptotic expansions to compare mean square errors or they used Monte Carlo sampling. Because the problem of estimation of error rates is most acute when the sample sizes are small, asymptotic expansions might not give results of direct interest for applications. Sorum (1971) wrote "... conclusions ... based on asymptotic MSEs, rather than on the distribution of the estimators for finite sample sizes, ... should perhaps not be taken as direct recommendations for the estimation of ... $[\alpha(\hat{\zeta})]$ in a practical situation".

Monte Carlo sampling studies have been done by several authors, but in none of those reviewed was UMSE the criterion for comparison. According to Sedransk and Okamoto (1971), "... the criterion of unconditional mean square error ... both seems a more meaningful measure of performance and provides clearer discrimination among the estimators ...." While the mean absolute error (1.10) is probably a good criterion, it is of at least equal interest to compare estimators using unconditional mean square error (1.8).

Monte Carlo sampling can be used in nearly any situation, but a different technique can be applied when two simplifications are adopted. Glick (1978) wrote, "I believe that estimator performance relative to one another will be qualitatively similar for univariate and multivariate distributions...." Another common simplification, employed by Sorum (1971, 1972 and 1973), Sedransk and Okamoto (1971), and others, is to assume $\underset{\sim}{\Sigma}$ known. The resubstitution and leave-one-out estimators in the univariate normal case with k=1 are unaffected by this assumption (since the linear discriminant function in this case is independent of both σ^2 and s^2), and the results of Sedransk and Okamoto (1971) presented in Section I.3.2 tend to show that the quality of the parametric estimators relative to one another is not greatly affected. Since the parametric estimators rely on an estimate of the variance, the assumption of known variance will probably decrease their UMSEs relative to the resubstitution and leave-one-out estimators. However, the most important comparisons in this study are within the two classes of estimators, not between them.

Within this simplified setting, where π_1 and π_2 refer to populations with univariate normal distributions having a common known variance, it is relatively simple to use numerical integration as well as Monte Carlo sampling to compare estimators. Numerical integration can be more accurate and efficient than Monte Carlo sampling; it is to be preferred whenever the problem allows and no analytical solution can be found. Sayre (1980) used numerical integration to evaluate $E[\alpha(\hat{\zeta})]$, but it does not appear to have been used to compare error rate estimators. The first purpose of this study is to show how numerical integration can be used to evaluate $E[\hat{\alpha}-\alpha(\hat{\zeta})]^2$ and to compare estimators using this method.

I.4.2 Evaluation of smoothed estimators

Glick (1978) brought up the following points: 1) nonparametric estimators are to be preferred in practical situations due to the large unstable bias of the parametric estimators under non-normality;

2) both the bias and the variance should be considered when comparing nonparametric estimators; and 3) the leave-one-out estimator reduces the bias of the resubstitution estimator at the cost of additional variance. Based on these points, Glick suggested a new class of estimators, smoothed counting estimators, aimed at reducing the variance of the resubstitution estimator.

The main purpose of this study is to evaluate this new class of estimators, comparing it with existing estimators on the basis of UMSE. Both types of smoothing mentioned by Glick will be considered; namely, the smooth slide estimator and the estimator smoothed by the function Φ. Since smoothing is designed merely to reduce variance, it should be most effective applied to a counting estimator with large variance but small bias. Therefore, Glick's idea of smoothing will be applied to the leave-one-out estimator as well as the resubstitution estimator.

I.4.3 Effects of unequal sample sizes, unequal prior probabilities and non-normal distributions

Consider the following possible application of discriminant analysis: let π_1 refer to a population of people free of disease A, π_2 to those people who have disease A, and x to the result of a screening test for the disease. If disease A is rare, the prior probabilities of belonging to π_1 and π_2 are not equal, and there is probably a larger sample from π_1 than from π_2. For example, if a screening test for ischemic heart disease is performed in an asymptomatic population, there will be fewer diseased people to give false negative results than there are healthy people to give false positives. Since the error rates show the sensitivity and specificity of the test, their estimation can be extremely important. Although this is a common situation in discriminant analysis, the effects of $n_1 \neq n_2$ and $k \neq 1$ (1.1) on the estimation of error rates have not been thoroughly studied.

It is also possible that the screening test results do not have a normal distribution. While the parametric estimators have been shown to be better than the nonparametric estimators under normality, it is important to determine which estimators are most robust. Therefore, the error rate estimators will also be compared under some forms of non-normality.

I.5 Outline of the Specific Analyses Performed

The primary purpose of this study is to evaluate smoothed error rate estimators in discriminant analysis under a variety of conditions, and to compare these estimators to the existing ones. In this section, the specific analyses performed to fulfill that purpose are outlined.

Section I.2 describes many error rate estimators; however, only a subset of these are considered in the analyses. Three parametric estimators are chosen: the plug-in estimator, or D method, which was the first estimator to be suggested and is the simplest of the parametric estimators; and the DS and O methods, which are similar to the D method, but which contain bias correction terms. The resubstitution estimator, or R method, and the leave-one-out estimator, or U method, are chosen since they are the most commonly used nonparametric methods. The smooth slide and normal smoothing are each applied to both the R and U methods; these four types of smoothed estimators are evaluated with each of the constants $b = \frac{1}{4}, \frac{1}{2}, 1, 1\frac{1}{2}, 2$.

In order to compare these estimators, a series of computer programs are run. In each program the expected value, variance and UMSE are tabulated for several values of Δ (1.3), the square root of the Mahalanobis distance between the populations.

The estimators are compared by two separate methods: numerical integration and Monte Carlo sampling. The numerical integration study, presented in Chapter II, is fairly limited, due to complications in the evaluation of the theoretical integrals. Only the univariate normal case with known variance is considered, and smoothing is applied only to the resubstitution estimator. Also, when the prior probabilities are not equal, the leave-one-out estimator is not considered at all. Training samples of size ten and twenty-five are chosen, since they are felt to be in the range of values for which the estimation of error rates is extremely important. The specific situations considered by the numerical integration program are detailed in Table 1.1.

The Monte Carlo sampling study, presented in Chapter III, is much more general. All of the estimators are evaluated in each run, both with and without the assumption of known variance. In addition to the situations covered by the numerical integration program (this overlap serves to validate the accuracy of both programs), the Monte Carlo sampling program is also used to explore the r-variate normal case, and some types of non-normality; see Table 1.2.

Some of the parameters of the model are held fixed in both of the programs considered here. These are $\underset{\sim}{\mu}_1' = (0\ 0 \ldots 0)$, $\underset{\sim}{\mu}_2' = (\Delta\ 0 \ldots 0)$ and $\underset{\sim}{\Sigma} = \underset{\sim}{I}$. This may be done without loss of generality since the discriminant function, $W(\underset{\sim}{X})$, is invariant to linear transformations of the observation vector (Dunn 1971), as are all of the error rate estimators considered in Section I.2. In the Monte Carlo sampling program, the values of Δ which are used are $\Delta = 0, \frac{1}{2}, 1, 1\frac{1}{2}, 2, 3$; the numerical integration program additionally uses $\Delta = 2\frac{1}{2}, 3\frac{1}{2}, 4, 4\frac{1}{2}$.

TABLE 1.1

PARAMETERS FOR EIGHT NUMERICAL INTEGRATION RUNS IN THE UNIVARIATE NORMAL CASE WITH $\sigma^2=1$ ASSUMED KNOWN, $\mu_1=0$, AND $\mu_2=0,\frac{1}{2},1,\ldots,4\frac{1}{2}$

Run #	n_1	n_2	c
1	10	10	0
2	10	25	0
3	25	10	0
4	25	25	0
5	5	25	1.4
6	10	20	1.4
7	25	5	-1.4
8	20	10	-1.4

TABLE 1.2

PARAMETERS FOR THE INITIAL FOURTEEN MONTE CARLO RUNS

WITH $\Delta=0,\frac{1}{2},1,1\frac{1}{2},2,3$

Run #	n_1	n_2	c	r	Distribution
1	10	10	0	1	Normal
2	25	25	0	1	Normal
3	10	20	1.4	1	Normal
4	20	10	-1.4	1	Normal
5	10	10	0	3	Normal
6	25	25	0	3	Normal
7	10	10	0	5	Normal
8	25	25	0	5	Normal
9	10	10	0	1	Uniform
10	25	25	0	1	Uniform
11	10	10	0	1	Exponential
12	25	25	0	1	Exponential
13	10	10	0	1	Double Exponential
14	25	25	0	1	Double Exponential

The three non-normal distributions considered are the uniform, exponential, and double exponential distributions, all in the univariate case. For comparability, each of these distributions is transformed in order to obtain $\mu_1=0$, $\mu_2=\Delta$, and $\sigma^2=1$. The density functions for the distributions as they are used in the Monte Carlo sampling program are given in Figure 1.1.

The two sets of programs defined here comprise the initial study designed to explore the smoothed error rate estimators and examine the factors which affect their quality relative to the other estimators. In Chapter IV, one specific smoothed estimator is suggested as an alternative to the existing estimators. This estimator, called the NS method, is compared to the existing estimators in all of the situations shown in Tables 1.1 and 1.2 plus some additional situations. Since sample size and dimensionality are deemed to be two very important factors, the programs are run with some other values of n_1, n_2 and r. The additional situations covered in Chapter IV are shown in Table 1.3.

Finally, in Chapter V, the use of the NS method is illustrated on a data set collected to study two-year survival following myocardial infarction. In that chapter a FORTRAN subroutine is described which may be used to evaluate the NS estimate for several different types of discriminant functions.

FIGURE 1.1

DENSITY FUNCTIONS FOR THREE NON-NORMAL DISTRIBUTIONS

USED IN THE MONTE CARLO SAMPLING PROGRAM

1) Uniform Distribution

$$f_1(x) = \begin{cases} 1/\sqrt{12} & -\sqrt{3} \leq x \leq \sqrt{3} \quad \text{if } x \text{ from } \pi_1 \\ 1/\sqrt{12}, & \Delta-\sqrt{3} \leq x \leq \Delta+\sqrt{3} \quad \text{if } x \text{ from } \pi_2 \end{cases}$$

2) Exponential Distribution

$$f_2(x) = \begin{cases} \exp[-(x+1)], & -1 \leq x < \infty \quad \text{if } x \text{ from } \pi_1 \\ \exp[-(x-\Delta+1)], & \Delta-1 \leq x < \infty \quad \text{if } x \text{ from } \pi_2 \end{cases}$$

3) Double Exponential Distribution

$$f_3(x) = \begin{cases} \exp(-|x|/\sqrt{2}) \;/\; \sqrt{2}, & -\infty < x < \infty \quad \text{if } x \text{ from } \pi_1 \\ \exp(-|x-\Delta|/\sqrt{2}) \;/\; \sqrt{2}, & -\infty < x < \infty \quad \text{if } x \text{ from } \pi_2 \end{cases}$$

TABLE 1.3

PARAMETERS FOR ADDITIONAL RUNS,

ALL ASSUMING NORMALITY AND c=0

n_1	n_2	r	Program
5	5	1	Numerical Integration
100	100	1	Numerical Integration
5	5	2	Monte Carlo
10	10	2	Monte Carlo
25	25	2	Monte Carlo
5	5	3	Monte Carlo

II. THE NUMERICAL INTEGRATION STUDY

II.1 An Expression for $E[\hat{\alpha}-\alpha(\hat{\zeta})]^2$ for Counting-Type Estimators under Normality

II.1.1 Theory and notation

Let π_1 and π_2 refer to two populations. The random variable, X, has the following distribution:

$$X \sim \begin{cases} N(\mu_1,\sigma^2) & \text{if X is from } \pi_1 \\ N(\mu_2,\sigma^2) & \text{if X is from } \pi_2, \end{cases}$$

where σ^2 is a known constant. Training samples of size n_1 and n_2 are drawn from the two populations and the sample means are evaluated. These have the following distributions:

$$\bar{X}_1 \sim N(\mu_1,\sigma^2/n_1)$$
$$\bar{X}_2 \sim N(\mu_2,\sigma^2/n_2).$$

The linear discriminant function is

$$W(X) = [X - \tfrac{1}{2}(\bar{X}_1+\bar{X}_2)]\ (\bar{X}_1-\bar{X}_2)\ /\ \sigma^2, \qquad (2.1)$$

and the decision rule is

$$\text{if } W(x) \begin{cases} > c & \text{assign x to } \pi_1 \\ \leq c & \text{assign x to } \pi_2 \end{cases}$$

where $c = \ln(k)$ may or may not be zero. The conditional actual error rate is the probability that a random observation from π_1 is misclassified by the decision rule (2.1), given $\bar{x}_1$ and $\bar{x}_2$. Assuming X is from π_1, this is

$$\alpha(\hat{\zeta}) = \Pr\{ W(X) \leq c \mid \bar{x}_1, \bar{x}_2 \}.$$

II.1.2 Conditioning on $\bar{x}_1$ and $\bar{x}_2$

In order to evaluate $E[\hat{\alpha}-\alpha(\hat{\zeta})]^2$, the first step is to condition on the sample means:

$$E[\hat{\alpha}-\alpha(\hat{\zeta})]^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\{ [\hat{\alpha}-\alpha(\hat{\zeta})]^2 \mid \bar{x}_1,\bar{x}_2 \}\, f_{\bar{X}_1}(\bar{x}_1)\, f_{\bar{X}_2}(\bar{x}_2)\, d\bar{x}_2\, d\bar{x}_1$$

$$\begin{aligned} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{ E[\alpha^2(\hat{\zeta}) \mid \bar{x}_1,\bar{x}_2] - 2E[\hat{\alpha}\alpha(\hat{\zeta}) \mid \bar{x}_1,\bar{x}_2] \\ &\quad + E[\hat{\alpha}^2 \mid \bar{x}_1,\bar{x}_2] \}\, f_{\bar{X}_1}(\bar{x}_1)\, f_{\bar{X}_2}(\bar{x}_2)\, d\bar{x}_2\, d\bar{x}_1. \end{aligned} \tag{2.2}$$

Conditioned on $\bar{x}_1$ and $\bar{x}_2$, $\alpha(\hat{\zeta})$ is a constant:

$$\begin{aligned} \alpha(\hat{\zeta}) &= \Pr\{ W(X) \leq c \mid \bar{x}_1,\bar{x}_2 \} \\ &= \Pr\{ [X-\tfrac{1}{2}(\bar{x}_1+\bar{x}_2)]\,(\bar{x}_1-\bar{x}_2) / \sigma^2 \leq c \} \\ &= \begin{cases} \Phi\{ -[\mu_1-\tfrac{1}{2}(\bar{x}_1+\bar{x}_2)]/\sigma + c\sigma/(\bar{x}_1-\bar{x}_2) \} & \text{if } \bar{x}_1 > \bar{x}_2 \\ 1 - \Phi\{ -[\mu_1-\tfrac{1}{2}(\bar{x}_1+\bar{x}_2)]/\sigma + c\sigma/(\bar{x}_1-\bar{x}_2) \} & \text{if } \bar{x}_1 < \bar{x}_2. \end{cases} \end{aligned}$$

In order to solve (2.2) it is still necessary to obtain expressions for $E(\hat{\alpha} \mid \bar{x}_1,\bar{x}_2)$ and $E(\hat{\alpha}^2 \mid \bar{x}_1,\bar{x}_2)$.

II.1.3 The general counting-type estimator

Let X_i, $i = 1,\ldots,n_1$, denote the members of the training sample from π_1. In addition, let U and V be two unique random members of this group. The general counting-type estimator can be written as

$$\hat{\alpha} = n_1^{-1} \sum_{i=1}^{n_1} h'(X_i).$$

Define $h(X) = h'(X) \mid \bar{x}_1, \bar{x}_2$; the expected values of $\hat{\alpha}$ and $\hat{\alpha}^2$ conditioned on $\bar{x}_1$ and $\bar{x}_2$ are

$$E(\hat{\alpha} \mid \bar{x}_1, \bar{x}_2) = E[\, n_1^{-1} \sum_{i=1}^{n_1} h(X_i)\,] = E[h(U)] \tag{2.3}$$

and

$$\begin{aligned} E(\hat{\alpha}^2 \mid \bar{x}_1, \bar{x}_2) &= E[\, n_1^{-2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} h(X_i)\, h(X_j)\,] \\ &= n_1^{-2}\, E[\, \sum_{i=1}^{n_1} h^2(X_i) + \sum_{\substack{i=1 \\ i\neq j}}^{n_1} \sum_{j=1}^{n_1} h(X_i)\, h(X_j)\,] \\ &= n_1^{-1}\, E[h^2(U)] + (n_1-1)n_1^{-1}\, E[h(U)h(V)]. \end{aligned} \tag{2.4}$$

It can easily be shown that the joint distribution of U and V, given $\bar{x}_1$ and $\bar{x}_2$, is:

$$\begin{pmatrix} U \\ V \end{pmatrix} \Big| \; \bar{x}_1, \bar{x}_2 \sim N_2 \left[\begin{pmatrix} \bar{x}_1 \\ \bar{x}_1 \end{pmatrix}, \frac{(n_1-1)\sigma^2}{n_1} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right] \tag{2.5}$$

where $\rho = -(n_1-1)^{-1}$. Therefore

$$E[h(U)] = \int_{-\infty}^{\infty} h(u)\; [n_1/2\pi\sigma^2(n_1-1)]^{\frac{1}{2}} \exp[-n_1(u-\bar{x}_1)^2/2\sigma^2(n_1-1)]\; du.$$

Letting $Y = [n_1/\sigma^2(n_1-1)]^{\frac{1}{2}}\,(U-\bar{x}_1)$,

$$E[h(U)] = \int_{-\infty}^{\infty} h\{\, [\sigma^2(n_1-1)/n_1]^{\frac{1}{2}} y + \bar{x}_1 \,\}\; \phi(y)\; dy, \tag{2.6}$$

where $\phi(t) = (2\pi)^{-\frac{1}{2}} \exp(-t^2/2)$. Similarly,

$$E[h^2(U)] = \int_{-\infty}^{\infty} h^2\{\, [\sigma^2(n_1-1)/n_1]^{\frac{1}{2}} y + \bar{x}_1 \,\}\; \phi(y)\; dy. \tag{2.7}$$

The expression for $E[h(U)h(V)]$ is somewhat more complicated. Using (2.5),

$$E[h(U)h(V)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(u)h(v)\ [n_1/2\pi\sigma^2(n_1-1)]\ (1-\rho^2)^{-\frac{1}{2}}$$
$$\exp\left\{-\frac{(u-\bar{x}_1)^2 + (v-\bar{x}_1)^2 - 2\rho(u-\bar{x}_1)(v-\bar{x}_1)}{2\sigma^2(n_1-1)(1-\rho^2)/n_1}\right\} du\ dv.$$

Letting $Y = [n_1/\sigma^2(n_1-1)]^{\frac{1}{2}}\ (U-\bar{x}_1)$ and $Z = [n_1/\sigma^2(n_1-1)]^{\frac{1}{2}}\ (V-\bar{x}_1)$, this expression reduces to

$$E[h(U)h(V)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h\{\ [\sigma^2(n_1-1)/n_1]^{\frac{1}{2}}y + \bar{x}_1\ \}$$
$$h\{\ [\sigma^2(n_1-1)/n_1]^{\frac{1}{2}}z + \bar{x}_1\ \}\ g(y,z;\rho)\ dz\ dy, \qquad (2.8)$$

where $g(t_1,t_2;\rho) = [2\pi(1-\rho^2)^{\frac{1}{2}}]^{-1} \exp[\ -(t_1^2 + t_2^2 - 2\rho t_1 t_2)\ /\ 2(1-\rho^2)\ ]$.

Therefore, by substituting (2.6) into (2.3), and (2.7) and (2.8) into (2.4), expressions have been found for $E(\hat{\alpha} \mid \bar{x}_1,\bar{x}_2)$ and $E(\hat{\alpha}^2 \mid \bar{x}_1,\bar{x}_2)$ in terms of integrals involving the function h.

II.1.4 Expressions for h for the smoothed estimators

The function $h(\underset{\sim}{x})$ will now be defined for each of the four types of smoothed estimators. Although defined by Glick (1978) only in the univariate case, they are here extended to the multivariate case.

Define the following:

$$m^2 = \begin{cases} (\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2)'\ \underset{\sim}{\Sigma}^{-1}\ (\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2) & \text{if } \underset{\sim}{\Sigma} \text{ known} \\ (\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2)'\ \underset{\sim}{S}^{-1}\ (\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2) & \text{if } \underset{\sim}{\Sigma} \text{ unknown} \end{cases}$$

$$w(\underset{\sim}{x}) = \begin{cases} [\underset{\sim}{x}-\tfrac{1}{2}(\bar{\underset{\sim}{x}}_1+\bar{\underset{\sim}{x}}_2)]'\ \underset{\sim}{\Sigma}^{-1}\ (\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2) & \text{if } \underset{\sim}{\Sigma} \text{ known} \\ [\underset{\sim}{x}-\tfrac{1}{2}(\bar{\underset{\sim}{x}}_1+\bar{\underset{\sim}{x}}_2)]'\ \underset{\sim}{S}^{-1}\ (\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2) & \text{if } \underset{\sim}{\Sigma} \text{ unknown.} \end{cases}$$

Smooth slide applied to resubstitution (RS method):

$$h^{RS}(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } w(\underset{\sim}{x})-c \le -bm \\ 0 & \text{if } w(\underset{\sim}{x})-c \ge bm \\ [w(\underset{\sim}{x})-c+bm]\ /\ 2bm & \text{otherwise.} \end{cases} \qquad (2.9)$$

Smooth slide applied to leave-one-out (US method):

Let $\bar{\underset{\sim}{x}}_1^*$ be the sample mean of the n_1-1 observations from π_1 when $\underset{\sim}{x}$ is omitted, $\bar{\underset{\sim}{x}}_1^* = (n_1-1)^{-1}\ (n_1\bar{\underset{\sim}{x}}_1-\underset{\sim}{x})$. Therefore,

$$h^{US}(\underset{\sim}{x};\bar{\underset{\sim}{x}}_1) = h^{RS}(\underset{\sim}{x};\bar{\underset{\sim}{x}}_1^*).$$

Normal smoothing applied to resubstitution (RN method):

$$h^{RN}(\underset{\sim}{x}) = \Phi\{\ [c-w(\underset{\sim}{x})]\ /\ bm\ \}. \qquad (2.10)$$

Normal smoothing applied to leave-one-out (UN method):

$$h^{UN}(\underset{\sim}{x};\bar{\underset{\sim}{x}}_1) = h^{RN}(\underset{\sim}{x};\bar{\underset{\sim}{x}}_1^*).$$

Note that with $b = 0$, both the RS and RN methods reduce to the R method; similarly, the US and UN methods reduce to the U method.

II.1.5 Expressions for integrals involving h(x)

The expressions for $E[h(U)]$ (2.6), $E[h^2(U)]$ (2.7), and $E[h(U)h(V)]$ (2.8) for the specific functions h^{RS} and h^{RN} defined in Section II.1.4 seem to require numerical integration in order to be evaluated. In this section, these formulas are written in terms of functions which are evaluated by readily available computer packages, eliminating the need for numerical integration. The following derivations all apply to resubstitution with $\bar{x}_1 > \bar{x}_2$. The expressions for resubstitution with $\bar{x}_1 < \bar{x}_2$ follow directly from these. Similar derivations for the leave-one-out estimator with $b = 0$ and $c = 0$ were performed, but are not presented here.

First let

$$d = [n_1/(n_1-1)]^{\frac{1}{2}}\ [c\sigma/(\bar{x}_1-\bar{x}_2) - (\bar{x}_1-\bar{x}_2)/2\sigma]$$

and

$$\beta = [n_1/(n_1-1)]^{\frac{1}{2}}\ b.$$

<u>Case 1:</u> $b = 0$

The function $h^{RS}\{\ [\sigma^2(n_1-1)/n_1]^{\frac{1}{2}}y + \bar{x}_1\ \}$ is of interest here. From (2.9) it can be shown that

$$h^{RS}\{\ [\sigma^2(n_1-1)/n_1]^{\frac{1}{2}}y + \bar{x}_1\ \} = \begin{cases} 1 & \text{if } y \le d \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

Using (2.11) in (2.6) and (2.7) gives

$$E[h(U)] = E[h^2(U)] = \int_{-\infty}^{d} \phi(y)\ dy = \Phi(d).$$

Using (2.11) in (2.8) gives

$$E[h(U)h(V)] = \int_{-\infty}^{d} \int_{-\infty}^{d} g(y,z;\rho)\ dz\ dy = G(d,d;\rho).$$

The function $G(d,d;\rho)$ is the cumulative distribution function for a standard bivariate normal random variable with correlation ρ. It, and the function $\Phi(d)$, are both available in the IMSL library (1977).

<u>Case 2:</u> $b > 0$, normal smoothing

From (2.10), it can be shown that

$$h^{NS}\{\ [\sigma^2(n_1-1)/n_1]^{\frac{1}{2}}y + \bar{x}_1\ \} = \Phi\{(d-y)/\beta\}. \tag{2.12}$$

Using (2.12) in (2.6) gives

$$E[h(U)] = \int_{-\infty}^{\infty} \Phi\{(d-y)/\beta\}\ \phi(y)\ dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{d-y}{\beta}} \phi(t)\ \phi(y)\ dt\ dy$$

$$= Pr\{\ T < (d-Y)/\beta\ \} = Pr\{\ T^* < d/\beta\ \},$$

where $T^* = T + Y/\beta$ is distributed $N[0,(1+\beta^2)/\beta^2]$.

Therefore,

$$E[h(U)] = \Phi\{d/(1+\beta^2)^{\frac{1}{2}}\}$$

Using (2.12) in (2.7) gives

$$E[h^2(U)] = \int_{-\infty}^{\infty} \Phi^2\{(d-y)/\beta\}\ \phi(y)\ dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{d-y}{\beta}} \int_{-\infty}^{\frac{d-y}{\beta}} \phi(s)\ \phi(t)\ \phi(y)\ ds\ dt\ dy$$

$$= \Pr\{\ S < (d-Y)/\beta,\ T < (d-Y)/\beta\ \} = \Pr\{\ S^* < d/\beta,\ T^* < d/\beta\ \},$$

where $S^* = S + Y/\beta$ and $T^* = T + Y/\beta$ are distributed

$$N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \frac{1+\beta^2}{\beta^2} \begin{pmatrix} 1 & (1+\beta^2)^{-1} \\ (1+\beta^2)^{-1} & 1 \end{pmatrix} \right].$$

Therefore,

$$E[h^2(U)] = G\{\ d/(1+\beta^2)^{\frac{1}{2}},\ d/(1+\beta^2)^{\frac{1}{2}};\ 1/(1+\beta^2)\ \}.$$

Using (2.12) in (2.8) gives

$$E[h(U)h(V)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi\{(d-y)/\beta\}\ \Phi\{(d-z)/\beta\}\ g(y,z;\rho)\ dy\ dz$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{d-y}{\beta}} \int_{-\infty}^{\frac{d-z}{\beta}} \phi(s)\ \phi(t)\ g(y,z;\rho)\ ds\ dt\ dy\ dz$$

$$= \Pr\{\ S < (d-Z)/\beta\ ,\ T < (d-Y)/\beta\ \}$$

$$= \Pr\{\ S^* < d/\beta\ ,\ T^* < d/\beta\ \}$$

where $S^* = S + Z/\beta$ and $T^* = T + Y/\beta$ are distributed

$$N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \frac{1+\beta^2}{\beta^2} \begin{pmatrix} 1 & \rho(1+\beta^2)^{-1} \\ \rho(1+\beta^2)^{-1} & 1 \end{pmatrix} \right].$$

Therefore

$$E[h(U)h(V)] = G\{\ d/(1+\beta^2)^{\frac{1}{2}},\ d/(1+\beta^2)^{\frac{1}{2}};\ \rho/(1+\beta^2)\ \}.$$

<u>Case 3:</u> b > 0, smooth slide

From (2.9) it can be shown that

$$h^{RS}\{\ [\sigma^2(n_1-1)/n_1]^{\frac{1}{2}}y + \bar{x}_1\ \} = \begin{cases} 1 & \text{if } y \le d-\beta \\ (d+\beta-y)/2\beta & \text{if } d-\beta < y < d+\beta \\ 0 & \text{if } d+\beta \le y. \end{cases} \quad (2.13)$$

Using (2.13) in (2.6) gives

$$E[h(U)] = \int_{-\infty}^{d-\beta} \phi(y)\ dy + \int_{d-\beta}^{d+\beta} [(d+\beta-y)/2\beta]\ \phi(y)\ dy$$

$$= \Phi(d-\beta) + [(d+\beta)/2\beta] \int_{d-\beta}^{d+\beta} \phi(y)\ dy + [2\beta]^{-1} \int_{d-\beta}^{d+\beta} \phi(y)\ (-ydy)$$

$$= \Phi(d-\beta) + [(d+\beta)/2\beta]\ [\Phi(d+\beta) - \Phi(d-\beta)] + [2\beta]^{-1}\ [\phi(d+\beta) - \phi(d-\beta)]$$

$$= [(d+\beta)\ \Phi(d+\beta) - (d-\beta)\ \Phi(d-\beta) + \phi(d+\beta) - \phi(d-\beta)]\ /\ 2\beta.$$

Using (2.13) in (2.7) gives

$$E[h^2(U)] = \int_{-\infty}^{d-\beta} \phi(y)\ dy + \int_{d-\beta}^{d+\beta} [(d+\beta-y)/2\beta]^2\ \phi(y)\ dy$$

$$= \Phi(d-\beta) + [(d+\beta)/2\beta]^2 \int_{d-\beta}^{d+\beta} \phi(y)\ dy + [(d+\beta)/2\beta^2] \int_{d-\beta}^{d+\beta} \phi(y)\ (-ydy) + (2\beta)^{-2} \int_{d-\beta}^{d+\beta} (2\pi)^{-\frac{1}{2}} \exp(-y^2/2)\ y^2\ dy$$

$$= \Phi(d-\beta) + [(d+\beta)/2\beta]^2 [\Phi(d+\beta) - \Phi(d-\beta)]$$

$$+ [(d+\beta)/2\beta^2] [\phi(d+\beta) - \phi(d-\beta)]$$

$$+ (2\beta)^{-2} \{ (d-\beta)(2\pi)^{-\frac{1}{2}} \exp[-(d-\beta)^2/2]$$

$$- (d+\beta)(2\pi)^{-\frac{1}{2}} \exp[-(d+\beta)^2/2] + \int_{d-\beta}^{d+\beta} \phi(y)\, dy \}$$

$$= \{ [(d+\beta)^2+1] \Phi(d+\beta) + [4\beta^2 - (d+\beta)^2 - 1] \Phi(d-\beta)$$

$$+ (d+\beta) \phi(d+\beta) - (d+3\beta) \phi(d-\beta) \} / 4\beta^2.$$

Using (2.13) in (2.8) gives

$$E[h(U)h(V)] = \int_{-\infty}^{d-\beta} \int_{-\infty}^{d-\beta} g(y,z;\rho)\, dy\, dz$$

$$+ \int_{-\infty}^{d-\beta} \int_{d-\beta}^{d+\beta} [(d+\beta-y)/2\beta]\; g(y,z;\rho)\, dy\, dz$$

$$+ \int_{d-\beta}^{d+\beta} \int_{-\infty}^{d-\beta} [(d+\beta-z)/2\beta]\; g(y,z;\rho)\, dy\, dz$$

$$+ \int_{d-\beta}^{d+\beta} \int_{d-\beta}^{d+\beta} [(d+\beta-y)/2\beta]\; [(d+\beta-z)/2\beta]\; g(y,z;\rho)\, dy\, dz.$$

This derivation proceeds in a fashion similar to those just presented. However, due to its length, only the result is presented here. The entire derivation is presented in Appendix 1. The result is

$$E[h(U)h(V)] = \{ [(d+\beta)^2+\rho]\, G(d+\beta,d+\beta;\rho) - 2[(d+\beta)(d-\beta)+\rho]\, G(d-\beta,d+\beta;\rho)$$

$$+ [(d-\beta)^2+\rho]\, G(d-\beta,d-\beta;\rho) + (1-\rho^2)\, g(d+\beta,d+\beta;\rho)$$

$$- 2(1-\rho^2)\, g(d-\beta,d+\beta;\rho) + (1-\rho^2)\, g(d-\beta,d-\beta;\rho)$$

$$+ 2(d+\beta)\, \phi(d+\beta)\, \Phi[(d+\beta-\rho d-\rho\beta)\, (1-\rho^2)^{-\frac{1}{2}}]$$

$$- 2(d-\beta)\, \phi(d+\beta)\, \Phi[(d-\beta-\rho d-\rho\beta)\, (1-\rho^2)^{-\frac{1}{2}}]$$

$$- 2(d+\beta)\, \phi(d-\beta)\, \Phi[(d+\beta-\rho d+\rho\beta)\, (1-\rho^2)^{-\frac{1}{2}}]$$

$$+ 2(d-\beta)\, \phi(d-\beta)\, \Phi[(d-\beta-\rho d+\rho\beta)\, (1-\rho^2)^{-\frac{1}{2}}] \} / 4\beta^2.$$

II.1.6 Summary

The purpose here is to evaluate the unconditional mean square error, or $E[\hat{\alpha} - \alpha(\hat{\zeta})]^2$, for any real value of c for several counting-type estimators in the univariate normal case with σ^2 known. The first step is to choose particular values of $\bar{x}_1$ and $\bar{x}_2$ and to calculate $\alpha(\hat{\zeta})$ (Section II.1.2); and $E[h(U)]$, $E[h^2(U)]$, and $E[h(U)h(V)]$ (Section II.1.5). Combine these to obtain:

$$E\{ [\hat{\alpha} - \alpha(\hat{\zeta})]^2 \mid \bar{x}_1,\bar{x}_2 \} = [\alpha(\hat{\zeta}) \mid \bar{x}_1,\bar{x}_2]^2 - 2[\alpha(\hat{\zeta}) \mid \bar{x}_1,\bar{x}_2]\, E[h(U)] + n_1^{-1}\, E[h^2(U)] + (n_1-1)n_1^{-1}\, E[h(U)h(V)].$$

All of these terms can be evaluated using programs available in the IMSL library.

The next step is to integrate over the distributions of $\bar{X}_1$ and $\bar{X}_2$, using the following relationship:

$$\begin{aligned} E[\hat{\alpha} - \alpha(\hat{\zeta})] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\{ [\hat{\alpha} - \alpha(\hat{\zeta})]^2 \mid \bar{x}_1,\bar{x}_2 \}\, f_{\bar{X}_1}(\bar{x}_1)\, f_{\bar{X}_2}(\bar{x}_2)\, d\bar{x}_1\, d\bar{x}_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\{ [\hat{\alpha} - \alpha(\hat{\zeta})]^2 \mid \bar{x}_1,\bar{x}_2 \}\, (n_1 n_2)^{\frac{1}{2}}\, (2\pi\sigma^2)^{-1} \\ &\qquad \exp[-n_1(\bar{x}_1-\mu_1)^2/2\sigma^2]\, \exp[-n_2(\bar{x}_2-\mu_2)^2/2\sigma^2]\, d\bar{x}_1\, d\bar{x}_2 \\ &= \pi^{-1} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\{ [\hat{\alpha} - \alpha(\hat{\zeta})]^2 \mid \bar{x}_1,\bar{x}_2 \}\, e^{-y_1^2-y_2^2}\, dy_1\, dy_2 \end{aligned}$$

where $Y_1 = (n_1/2\sigma^2)^{\frac{1}{2}}(\bar{X}_1-\mu_1)$ and $Y_2 = (n_2/2\sigma^2)^{\frac{1}{2}}(\bar{X}_2-\mu_2)$. This equation can be solved fairly easily using numerical integration.

The numerical integration technique used here is Simpson's rule (Carnahan, Luther and Wilks 1969, p. 73). While numerical integration in two dimensions requires considerably more computer resources than in one dimension, it can still be done easily and accurately.

Two methods of validating the accuracy of this numerical integration technique are considered. The first involves comparison of the expected values for the resubstitution, leave-one-out, and plug-in estimators calculated by numerical integration with those using exact formulas in the literature. The other involves comparisons of unconditional mean square errors calculated by numerical integration with those estimated by Monte Carlo sampling.

II.2 Results of the Numerical Integration Study

The program which performed the numerical integrations, written in FORTRAN, evaluates the unconditional mean square error (UMSE), mean, and variance for each of fifteen estimators: resubstitution; leave-one-out (only assuming $c = 0$); normal smoothing applied to resubstitution (RN method), with five values of the constant, b; smooth slide applied to resubstitution (RS method), also with five values of b; plug-in; the DS method; and the O method.

In addition the program evaluates six other values, all of which are derived from results in the literature and assume $c = 0$. These are: the expected actual error rate, $E[\alpha(\hat{\zeta})]$ (1.7); the expected values of the resubstitution estimator (1.13), plug-in estimator (1.12), and leave-one-out estimator (Hills 1966); and the UMSEs for the resubstitution and leave-one-out estimators based on asymptotic expansions (Sedransk and Okamoto 1971). None of these values require numerical integration for their evaluation.

In each application of the numerical integration program with $c = 0$ (see Table 1.1), the expected values for the three error rate estimators evaluated in this way were compared with the same values evaluated by numerical integration. In no instance was the difference greater than .001, suggesting that the results evaluated by numerical integration are accurate to three decimal places.

Numerical integration is only performed in the univariate normal case with variance assumed known. However, within this setting, the program allows considerable flexibility in the parameters which may be altered. Some of these parameters were held constant in all of the runs performed. These are $\mu_1 = 0$, $\sigma^2 = 1$, and the five levels of b (1/4, 1/2, 1, 3/2, 2). The parameters which varied among and within runs are μ_2, n_1, n_2, and c, as was shown in Table 1.1.

In the following sections several aspects of the results from these eight runs are discussed and evaluated.

II.2.1 The effect of smoothing on the resubstitution estimator

Smoothing appears to affect the resubstitution estimator in two ways: as the constant, b, increases from zero, the expected values of the two smoothed estimators monotonically approach one-half, and their variances monotonically approach zero. These effects are illustrated by the sample results presented in Tables 2.1 and 2.2. Since larger values of b clearly force $h^{RS}(\underset{\sim}{x})$ (2.9) and $h^{RN}(\underset{\sim}{x})$ (2.10) closer to one-half, the effects make intuitive sense.

Consider these two effects in terms of estimation of the expected actual error rate, a problem similar to that of estimation of the conditional actual error rate (Sorum 1973). The smoothed estimators always have smaller variance than resubstitution, and the larger the value of b the smaller it is. In this respect the smoothed estimators are always better than the R method. With respect to their expected values, however, the results are not so simple.

TABLE 2.1

RESULTS OF NUMERICAL INTEGRATION PROGRAM

WITH $n_1=n_2=10$, $\Delta=1$, AND $c=0$

Estimator	Mean	Variance	UMSE
R Method	.303	.015	.021
U Method	.318	.019	.024
D Method	.312	.006	.012
DS Method	.321	.005	.012
O Method	.317	.006	.012
RS Method			
b=0.25	.305	.012	.019
b=0.5	.310	.010	.016
b=1.0	.330	.006	.013
b=1.5	.356	.004	.012
b=2.0	.380	.003	.013
RN Method			
b=0.25	.308	.011	.017
b=0.5	.323	.007	.014
b=1.0	.359	.004	.012
b=1.5	.389	.002	.014
b=2.0	.410	.002	.017

Exact Results (not requiring numerical integration)

Expected Actual Error Rate	.3175
Expected Value of R Method	.3032
Expected Value of U Method	.3189
Expected Value of D Method	.3120
Asymptotic UMSE of R Method	.0212
Asymptotic UMSE of U Method	.0210

TABLE 2.2

RESULTS OF NUMERICAL INTEGRATION PROGRAM

WITH $n_1=5$, $n_2=25$, $\Delta=1$, AND $c=1.4$

Estimator	Mean	Variance	UMSE
R Method	.799	.047	.026
D Method	.786	.034	.014
DS Method	.791	.033	.012
O Method	.782	.034	.014
RS Method			
b=0.25	.798	.044	.022
b=0.5	.793	.041	.019
b=1.0	.777	.036	.016
b=1.5	.754	.033	.016
b=2.0	.730	.030	.019
RN Method			
b=0.25	.795	.042	.020
b=0.5	.783	.037	.017
b=1.0	.749	.032	.016
b=1.5	.715	.028	.021
b=2.0	.688	.024	.027

It is well known that the resubstitution estimator is optimistically biased relative to the expected actual error rate (Hills 1966). When the expected actual error rate is less than one-half (this is almost always true unless $c > 0$, in which case it may or may not be true), the smoothed estimators tend to become less optimistically biased as b increases to some point, b*. There the bias is zero, and thereafter the bias becomes increasingly pessimistic, approaching some upper bound. For example, in the situation covered by Table 2.1, b* appears to be approximately 0.7 for the RS method and 0.4 for the RN method.

The value of b* is different in every different situation. It seems to be large when the expected actual error rate is close to one-half or when the training sample sizes are small, and to be small when the expected actual error rate is far from one-half or the training sample sizes are large.

In those cases when the expected actual error rate is greater than one-half, two situations are possible. If the expected value of the resubstitution estimator is less than one-half, the optimistic bias of the smoothed estimators decreases toward some lower bound greater than zero. If the expected value of the resubstitution estimator is greater than one-half, the optimistic bias of the smoothed estimators increases to an upper bound as b increases. This latter situation is illustrated in Table 2.2. Although the expected actual error rate is not given, presumably it is greater than .799, the expected value of the R method.

Although the above discussion deals with estimation of the expected actual error rate, it is felt that the effects in terms of estmation of the conditional actual error rate are roughly similar. This is explored further in Section II.2.3, where the UMSEs of the estimators are analyzed.

II.2.2 Normal smoothing versus smooth slide

The discussion of Section II.2.1 applies equally to the normal smoothing (RN) and smooth slide (RS) estimators. Based on their expected values and variances there does not seem to be any evidence for the preference of one type of smoothing over the other.

Although both the RN and RS estimators are defined using a constant labelled b, this constant does not have the same interpretation for the two estimators. For a given value of b, the RN method seems to achieve more smoothing than the RS method; in other words the RN method has a smaller variance and an expected value closer to one-half (see Tables 2.1 and 2.2).

This does not mean, however, that the normal smoothing estimator is any better than the smooth slide estimator. The better type of smoothing is felt to be the one which achieves the greater variance reduction for a given change in expected value. Based on this criterion there does not seem to be any basis for choosing one type of smoothing over another.

II.2.3 A comparison of estimators when sample sizes and prior probabilities are equal

Figures 2.1-2.4 contain plots of UMSE versus Δ for a sample of four of the eight situations described in Table 1.1. Since the normal smoothing (RN) and smooth slide (RS) estimators are felt to be nearly equivalent, it was not deemed necessary to present both of them in these figures. The smooth slide estimators were arbitrarily chosen to be included. Figure 2.1 corresponds to a situation with equal sample sizes and prior probabilities, which is of interest in this section.

The most obvious feature of this plot is that the leave-one-out estimator (or U method) has by far the largest UMSE when Δ is near zero. As Δ increases the U method improves so that the UMSEs of the U and R methods become approximately equal. The UMSEs of the two estimators follow the same pattern when the training samples are of size ten, but they approach less rapidly as Δ increases.

The three parametric estimators (D, DS, and O methods) are very nearly equal in terms of UMSE, and, as would be expected under normality, all have considerably lower UMSE than the U and R methods.

The smooth slide estimators seem to compare well with the other estimators, but there are a few factors to consider. The value of b has a large effect on the UMSE of the estimator. When the populations are very close, the more smoothing the lower the UMSE; the smoothed estimators all have smaller UMSE than resubstitution and the larger the value of b the smaller the UMSE becomes. For moderate and large separations between the populations, however, the smoothed estimators have smaller UMSE than the R method for small values of b, but as b increases the UMSE of the smoothed estimator can increase dramatically. Therefore, the optimum choice of b depends upon the unknown separation between the populations.

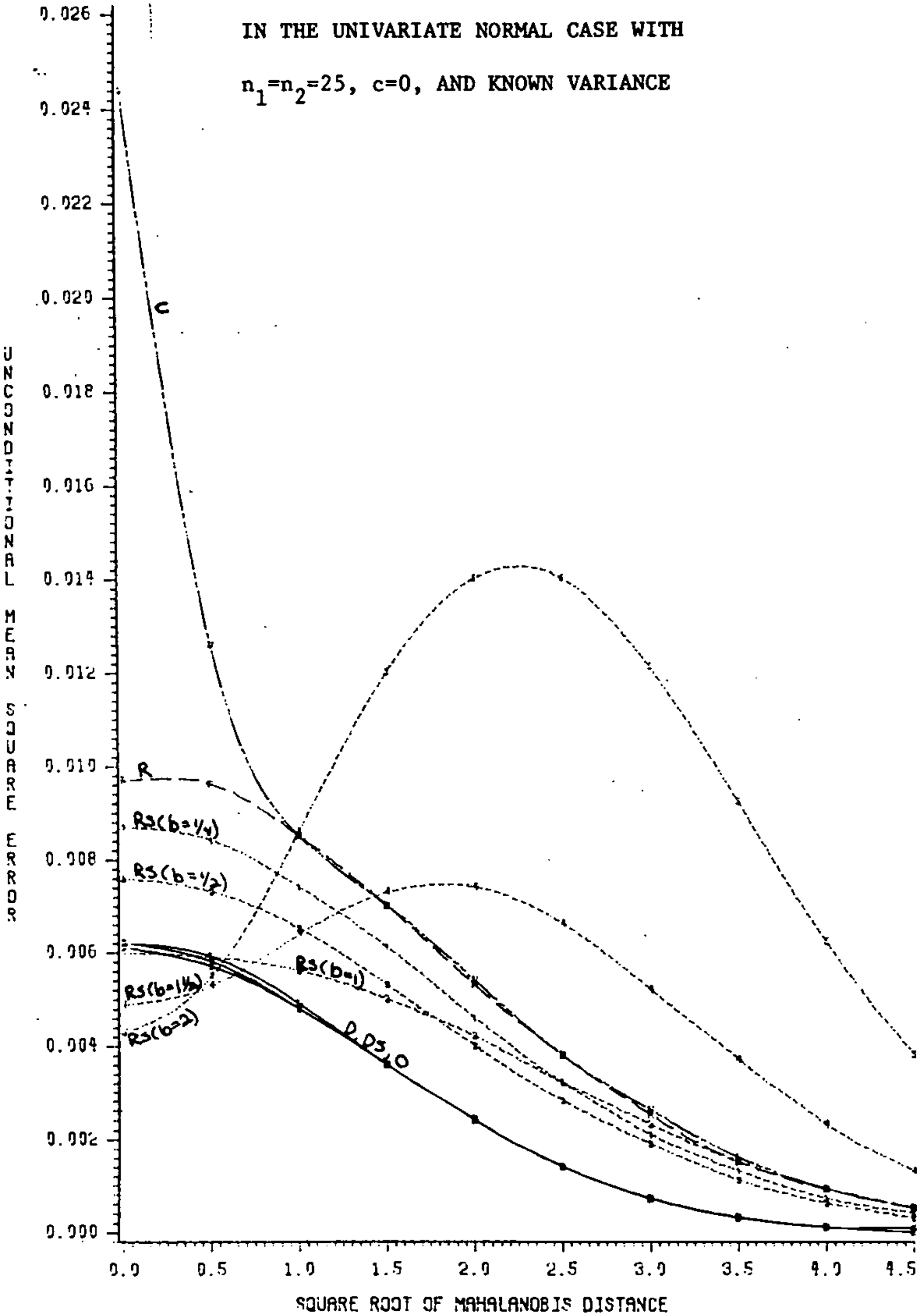
If this were the only factor influencing the choice of b, the solution would be fairly simple. For instance, in Figure 2.1, one could choose b equal to one-half and obtain an estimator which has

FIGURE 2.1

UMSE EVALUATED BY NUMERICAL INTEGRATION

IN THE UNIVARIATE NORMAL CASE WITH

$n_1=n_2=25$, $c=0$, AND KNOWN VARIANCE



UMSE smaller than those of the U and R methods and nearly as small as those of the parametric estimators. Unfortunately, the sample sizes also seem to influence the choice of b. When the training sample sizes are smaller, a larger value of b (between one and one and one-half for training samples of size ten) achieves the same result.

The results indicate that as sample size increases, the optimum value of b decreases, but no general relationship is apparent. In the next two sections, two other factors which might influence the choice of b are considered.

II.2.4 The effect of unequal sample sizes

Figures 2.2 and 2.3 correspond to situations identical to that in Figure 2.1 except for one of the training sample sizes. Figure 2.2, in which n_1=25 and n_2=10, is virtually identical in appearance to Figure 2.1, while Figure 2.3, in which n_1=10 and n_2=25, is very dissimilar. The conclusion to be drawn from these figures, and from other results not shown, seems to be that n_1 has a much larger effect than n_2 on the UMSEs of the error rate estimators.

It should be emphasized that the error of incorrectly classifying an observation from π_1 is the error rate under consideration. A more general conclusion is that, when considering the estimation of the error of misclassifying an observation from population i, the size of the training sample from population i is much more important than that from the other population.

II.2.5 The effect of unequal prior probabilities

Figure 2.4 corresponds to a situation in which the prior probability for population 1 is smaller than for population 2 (.20 to

FIGURE 2.2

UMSE EVALUATED BY NUMERICAL INTEGRATION

IN THE UNIVARIATE NORMAL CASE WITH

n_1=25, n_2=10, c=0, AND KNOWN VARIANCE

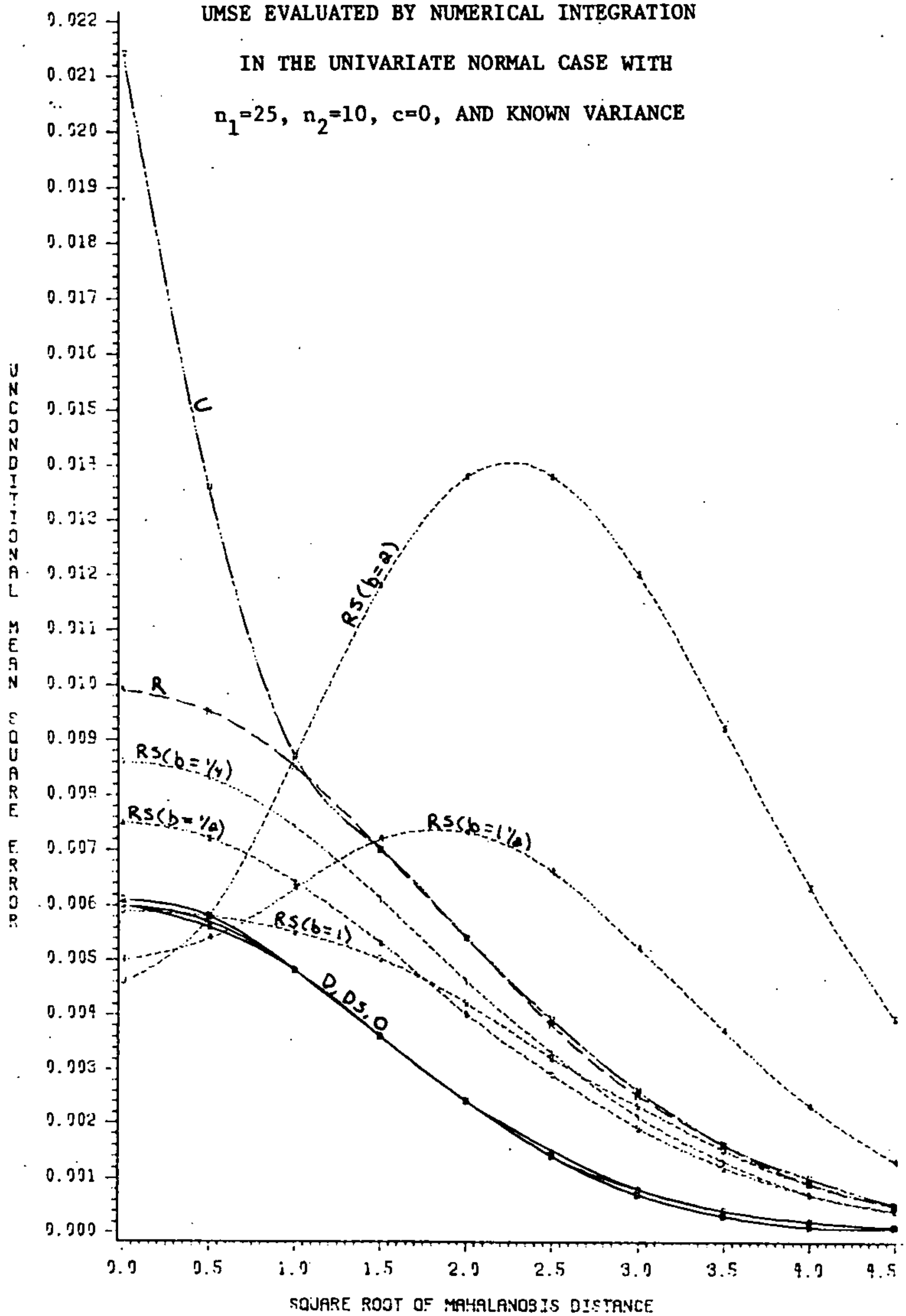
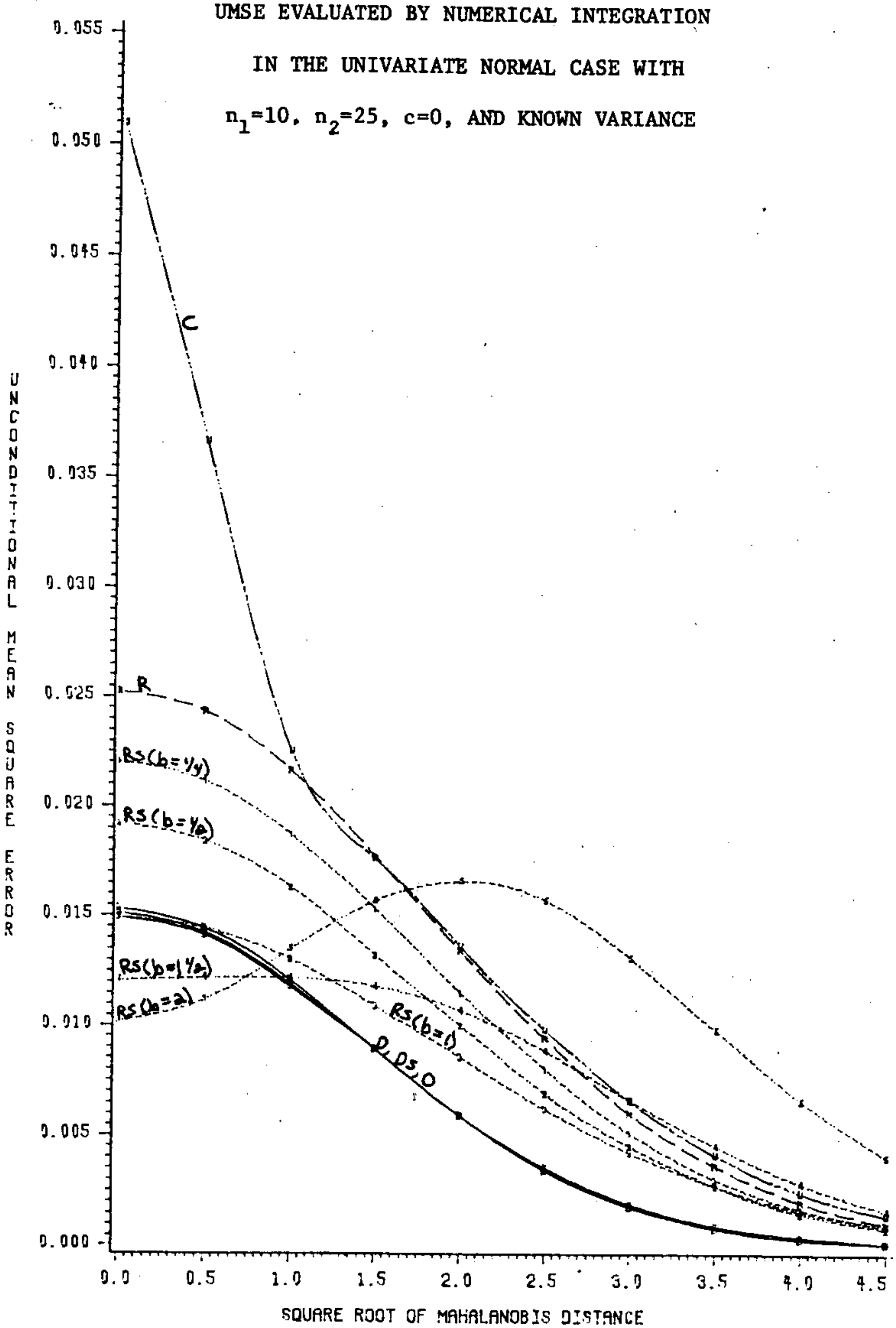


FIGURE 2.3

UMSE EVALUATED BY NUMERICAL INTEGRATION

IN THE UNIVARIATE NORMAL CASE WITH

n_1=10, n_2=25, c=0, AND KNOWN VARIANCE



.80). This figure appears to be very different from Figures 2.1 - 2.3, in which the prior probabilities are equal. However, the conclusions about the UMSEs of the estimators, relative to one another, to be drawn from it are virtually the same.

When the prior probabilities are equal, the UMSEs of the error rate estimators (with the exception of the smooth slide estimators with large b) are at a maximum when $\Delta=0$. When the prior probabilities are not equal, however, the maximum is achieved at some value of Δ greater than zero. For instance, in Figure 2.4, the UMSEs of all estimators (with the same exception as above) are maximized with $\Delta=2$.

At the value of Δ where the UMSEs are maximized, the values of the UMSEs of the error rate estimators relative to one another appear to be the same whether or not the prior probabilities are equal. Compare Figures 2.3 and 2.4, which both have $n_1=10$. The UMSEs of the estimators with $\Delta=0$ in Figure 2.3 are approximately equal to the same UMSEs with $\Delta=2$ in Figure 2.4. Similarly, the UMSEs with $\Delta=2$ in Figure 2.3 are approximately equal to those with $\Delta=3$ in Figure 2.4. The result here seems to be that inequality of prior probabilities is not an important factor in comparing the UMSEs of the error rate estimators.

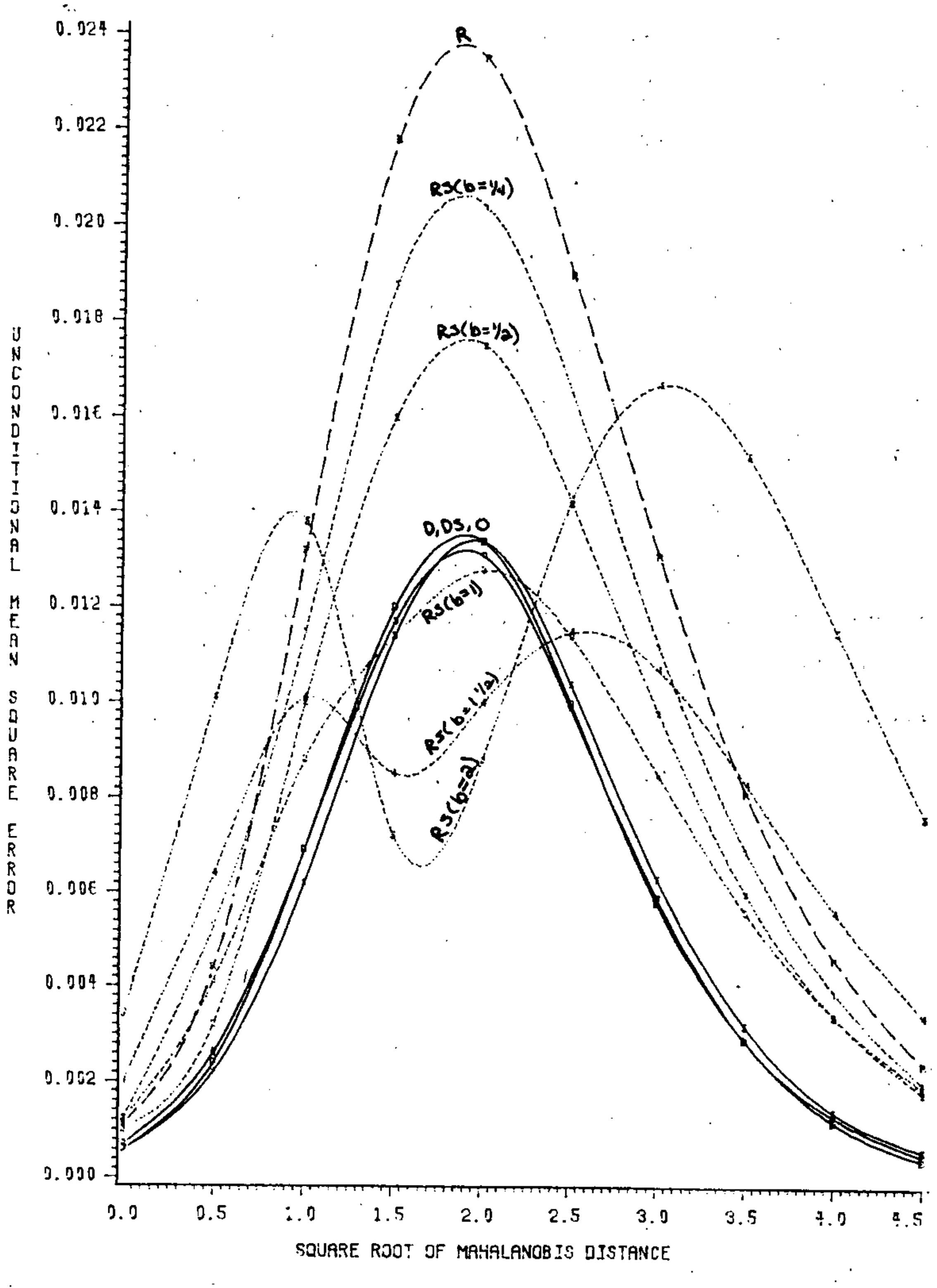
Note that the leave-one-out estimator is not presented in Figure 2.4 because this estimator is not evaluated by the numerical integration program with $c\neq0$.

II.2.6 Summary

Some of the main results learned from the numerical integration study about the univariate normal case with σ^2 known are as follows. With c=0, the U method never has UMSE much lower than the R method, but, if the populations are close, it can be much greater. The three

FIGURE 2.4

UMSE EVALUATED BY NUMERICAL INTEGRATION IN THE UNIVARIATE NORMAL CASE WITH $n_1=10$, $n_2=20$, $c=1.4$, AND KNOWN VARIANCE



parametric methods have very nearly equal UMSEs, which are considerably lower than those of the U and R methods. The UMSE of the smoothed estimators depends greatly on the choice of the constant b. At best, a smoothed estimator has UMSE much smaller than the U and R methods and not much larger than those of the parametric methods. The optimum choice of the constant depends upon the size of the training sample from π_1, but does not seem to depend much on the size of the sample from π_2 or on the prior probabilities.

The numerical integration study has explored the effects of several factors on the estimation of error rates, but it has omitted many others. A Monte Carlo sampling study, described in Chapter III, examines some of the other factors. In the univariate normal case the Monte Carlo sampling study considers the effect of the assumption of known variance on each of the estimators, and looks at smoothed versions of the leave-one-out estimator, with and without equal prior probabilities. This study also considers the effects of multivariate normal, and some non-normal, parent distributions.

III. THE MONTE CARLO SAMPLING STUDY

III.1 Properties of the Monte Carlo Program

The Monte Carlo sampling program allows much more flexibility than the numerical integration program. Along with all of the estimators included in the numerical integration program, the Monte Carlo program adds smoothed versions of the leave-one-out estimator, and no estimator is excluded due to unequal prior probabilities, as is done in the numerical integration program. In addition, all of the estimators in the Monte Carlo sampling program are evaluated both with and without the assumption of known variance. While the numerical integration program considers only the case of univariate normal parent distributions, the Monte Carlo program allows r-variate normal and some types of non-normal parents. The parameters $\underset{\sim}{\mu}_1=\underset{\sim}{0}$, $\underset{\sim}{\Sigma}=\underset{\sim}{I}$, and the five levels of smoothing (b = 1/4, 1/2, 1, 3/2, 2) were held fixed throughout all of the runs performed for this study, while the parameters $\underset{\sim}{\mu}_2$, n_1, n_2, c, r, and the class of parent distribution vary among and within runs.

In the remainder of this section, two aspects of the Monte Carlo sampling program are examined: the pseudorandom number generator, and the accuracy of the results. In Section III.2, the results of the Monte Carlo program are discussed.

III.1.1 The pseudorandom number generator

Uniform (0,1) pseudorandom numbers are generated in the Monte Carlo sampling program using the mixed multiplicative congruential method (Carnahan, Luther and Wilks 1969, p. 588). The general recursive relation for this method is

$$x_i = (ax_{i-1} + c) \bmod m,$$

where a,c, and m are integer constants, and the values $y_i = x_i/m$ are uniformly distributed between zero and one.

The maximum period for this sequence is m numbers, and this maximum is achieved provided the following three conditions are met:

i) c and m do not have a common divisor (3.1)

ii) a mod p = 1, where p is any prime factor of m

iii) a mod 4 = 1, if m is a multiple of 4

(Carnahan, Luther and Wilks 1971, p. 588).

The constants used in the program, a=509, c=886361, and $m=2^{22}=$ 4194304, were chosen for the following reasons: they meet the three criteria in (3.1); they reduce the upper bound for the autocorrelation between successive numbers, based on a formula given by Hammersley and Hanscomb (1967, p. 29); and they yield a period large enough to satisfy the requirements of the program. The maximum number of pseudorandom numbers generated by the program for any given set of parameters is 2,500,000, which is less than m.

Normal pseudorandom numbers, labelled z_i, are obtained from the uniform numbers by the method of Box and Muller (1958). Chen (1971) and Neave (1975) have cast some doubt about the quality of normal pseudorandom numbers obtained in the manner used here by the Monte Carlo sampling program; for this reason, some of the properties of these pseudorandom numbers are examined.

The distribution of the pseudorandom normal numbers is examined by calculating the first nine moments, $\overline{z^k}$, $k = 1,\ldots,9$. The results, based on a sample of 100,000 consecutive numbers and presented in Table 3.1, suggest that the second, and possibly fourth, moments of the pseudorandom numbers are less than for a true standard normal distribution. However, the differences are very small, and all of the other moments are close to their theoretical expected values. Based on this result, the pseudorandom numbers seem to have a very nearly normal distribution.

The autocorrelations of the normal pseudorandom numbers are also examined, based on the same sample of 100,000 consecutive numbers. The hypotheses considered are that the means of $z_j z_{j-i}$, $i = 1,\ldots,9$, are equal to zero. If the z_j act as independently distributed standard normal random numbers, then the mean of the 100,000-i values of $z_j z_{j-i}$, $\overline{zz_{-i}}$, would have an expected value of zero, and a standard error of approximately .00316. The results, presented in Table 3.2, indicate that autocorrelations are not a problem for these pseudorandom numbers.

III.1.2 Accuracy of the results

The Monte Carlo sampling program estimates the expected value, variance, and UMSE for each error rate estimator. In this section, the accuracy of these estimates is examined.

As a first step, expressions are obtained for the standard errors of the three estimates in terms of the number of repetitions in the

TABLE 3.1

MOMENTS OF 100,000 CONSECUTIVE PSEUDORANDOM NORMAL NUMBERS, WHERE $\overline{z^k}$ IS THE MEAN OF THE VALUES OF z_i^k, $i=1,\ldots,100{,}000$; AND $E(\overline{Z^k})$ AND s.e.$(\overline{Z^k})$ ARE THE EXPECTED VALUE AND STANDARD ERROR OF $\overline{Z^k}$, ASSUMING Z HAS A STANDARD NORMAL DISTRIBUTION

k	$\overline{z^k}$	$E(\overline{Z^k})$	s.e.$(\overline{Z^k})$	$[\overline{z^k}-E(\overline{Z^k})]/$s.e.$(\overline{Z^k})$
1	3.9×10^{-3}	0	3.2×10^{-3}	1.24
2	9.8×10^{-1}	1	4.5×10^{-3}	-4.13
3	6.5×10^{-3}	0	1.2×10^{-2}	0.53
4	2.9×10^{0}	3	3.1×10^{-2}	-2.06
5	-2.9×10^{-2}	0	9.7×10^{-2}	-0.30
6	1.5×10^{1}	15	3.2×10^{-1}	-1.56
7	-1.0×10^{0}	0	1.2×10^{0}	-0.86
8	9.8×10^{1}	96	4.5×10^{0}	0.55
9	-1.9×10^{1}	0	1.9×10^{1}	-1.02

TABLE 3.2

AUTOCORRELATIONS OF 100,000 CONSECUTIVE PSEUDORANDOM NORMAL NUMBERS, WHERE $\overline{zz_{-i}}$ IS THE MEAN OF THE VALUES OF $z_j z_{j-i}$, $j=i+1,\ldots,100{,}000$; $E(\overline{YZ})$ AND $s.e.(\overline{YZ})$ THE EXPECTED VALUE AND STANDARD ERROR OF $\overline{YZ}$, ASSUMING Y AND Z ARE INDEPENDENTLY DISTRIBUTED NORMAL RANDOM VARIABLES

i	$\overline{zz_{-i}} \times 10^3$	$[\overline{zz_{-i}} - E(\overline{YZ})]/s.e.(\overline{YZ})$
1	-4.20	-1.33
2	3.22	1.02
3	1.16	0.37
4	0.29	0.09
5	-1.46	-0.46
6	2.72	0.86
7	-4.49	-1.42
8	2.09	0.66
9	-0.20	-0.06

Monte Carlo sampling (n) and the variance of the error rate estimator. Let $\hat{\alpha}$ refer to an arbitrary error rate estimator, and $V(\hat{\alpha})$ to its variance. The first result is

$$\text{s.e.}(\bar{X}) = [V(\hat{\alpha})/n]^{\frac{1}{2}}. \tag{3.2}$$

Snedecor and Cochran (1967 p. 89) give the following:

$$\text{s.e.}(\hat{\sigma}^2) = V(\hat{\alpha})\ [2/(n-1) + \gamma/n]^{\frac{1}{2}},$$

where γ is the kurtosis of $\hat{\alpha}$, defined such that $(\gamma+3)\ V^2(\hat{\alpha}) = E[\hat{\alpha} - E(\hat{\alpha})]^4$. For normally distributed random variables γ is zero, and so as an approximation:

$$\text{s.e.}(\hat{\sigma}^2) \doteq V(\hat{\alpha})\ [2/(n-1)]^{\frac{1}{2}}. \tag{3.3}$$

It is more difficult to find the standard error of the unconditional mean square error (1.8). However, by assuming that $\alpha(\hat{\zeta})$ is a constant and that $\hat{\alpha}$ is normally distributed, the following is obtained:

$$\text{UMSE} = (n-1)\hat{\sigma}^2/n + [\bar{X} - \alpha(\hat{\zeta})]^2.$$

Under the above assumptions, $\bar{X} - \alpha(\hat{\zeta}) \sim N[\ \text{Bias},\ V(\hat{\alpha})/n\ ]$, where $\text{Bias} = E[\hat{\alpha} - \alpha(\hat{\zeta})]$. It can easily be shown that

$$\text{Var}[\bar{X} - \alpha(\hat{\zeta})]^2 = 2V^2(\hat{\alpha})/n^2 + 4V(\hat{\alpha})\text{Bias}^2/n$$

and

$$\text{Var}[(n-1)\hat{\sigma}^2/n] = 2(n-1)V^2(\hat{\alpha})/n^2.$$

The final result is therefore

$$\text{s.e.(UMSE)} \doteq [2V^2(\hat{\alpha})/n + 4V(\hat{\alpha})\text{Bias}^2/n]^{\frac{1}{2}}. \qquad (3.4)$$

In every Monte Carlo sampling run, n=5000 is used. Although $V(\hat{\alpha})$ is unknown and varies considerably with different estimators and in different situations, one can still explore the magnitudes of the standard errors for extreme and typical values of $V(\hat{\alpha})$.

The largest value of $V(\hat{\alpha})$ observed in Table 2.1 or 2.2 is .047 (R method in Table 2.2). Using this value in Equations 3.2 and 3.3, along with n=5000, yields s.e.($\bar{X}$) $\doteq$ .0031 and s.e.($\hat{\sigma}^2$) $\doteq$ .0009. The standard error of UMSE increases with Bias, but, if Bias is not very large, it is very nearly equal to s.e.($\hat{\sigma}^2$). For example, using Bias=.15, $V(\hat{\alpha})$=.047, and n=5000 in Equation 3.4 yields s.e.(UMSE) $\doteq$ .0013.

Based on all of the results of the numerical integration program, however, a typical value of $V(\hat{\alpha})$ seems to be closer to .01. Using $V(\hat{\alpha})$=.01, n=5000, and Bias=.15 in Equations 3.2-3.4 yields s.e.($\bar{X}$) $\doteq$ .0014, s.e.($\hat{\sigma}^2$) $\doteq$ .0002, and s.e.(UMSE) $\doteq$ .0005. These results indicate that while $\bar{X}$ may be accurate only to two decimal places, $\hat{\sigma}^2$ and UMSE usually have three digit accuracy.

There is some overlap between the situations covered by numerical integration (Table 1.1) and Monte Carlo sampling (Table 1.2). A comparison of the results of the two studies in these situations is another way of assessing the accuracy of the two programs. For instance, Table 2.2 can be compared with the first three columns of Table 3.3. When all of the results are compared, the above result is confirmed: the variances and UMSEs estimated by the two programs

TABLE 3.3

RESULTS OF THE MONTE CARLO SAMPLING PROGRAM

WITH UNIVARIATE NORMAL PARENT DISTRIBUTIONS,

$n_1=n_2=10$, $\Delta=1$, AND $c=0$

	Known Variance			Unknown Variance		
Estimator	Mean	Variance	UMSE	Mean	Variance	UMSE
R Method	.302	.015	.022	.302	.015	.022
U Method	.318	.019	.024	.318	.019	.024
D Method	.312	.006	.012	.305	.007	.013
DS Method	.321	.005	.012	.314	.006	.013
O Method	.316	.006	.012	.309	.007	.013
RS Method						
b=0.25	.305	.012	.019	.304	.012	.019
b=0.5	.311	.010	.016	.310	.010	.017
b=1.0	.330	.006	.013	.327	.007	.013
b=1.5	.356	.004	.012	.351	.005	.012
b=2.0	.380	.003	.013	.375	.003	.013
RN Method						
b=0.25	.309	.011	.017	.308	.011	.017
b=0.5	.323	.007	.014	.321	.008	.014
b=1.0	.359	.004	.012	.355	.005	.013
b=1.5	.389	.002	.014	.385	.003	.014
b=2.0	.411	.002	.017	.407	.002	.017
US Method						
b=0.25	.320	.016	.021	.320	.016	.021
b=0.5	.325	.014	.019	.325	.014	.019
b=1.0	.342	.009	.015	.342	.010	.016
b=1.5	.364	.006	.014	.363	.007	.014
b=2.0	.386	.004	.015	.384	.005	.015
UN Method						
b=0.25	.323	.014	.019	.323	.015	.019
b=0.5	.336	.011	.016	.336	.011	.016
b=1.0	.368	.006	.014	.366	.007	.014
b=1.5	.395	.004	.015	.393	.004	.016
b=2.0	.415	.002	.018	.412	.003	.018

always agree to three decimal places, while the means sometimes agree to only two decimal places.

III.2 Results of the Monte Carlo Sampling Study

III.2.1 Smoothing applied to the leave-one-out estimator

The idea of smoothing was originally suggested in order to reduce the variance of the resubstitution estimator. While Glick (1977) hoped that smoothing would have the beneficial side effect of reducing the resubstitution estimator's bias ("The smoothed estimator ... not only will reduce variance, but also is likely to lower the optimistic bias of the [resubstitution estimator]."), variance reduction is the real goal. It seems reasonable, therefore, that smoothing will be of most benefit to a counting estimator which has large variance and small bias, such as the leave-one-out estimator. For this reason, smoothed versions of the leave-one-out estimator were evaluated in the Monte Carlo sampling program, and these results are examined in this section.

Smoothing appears to have the same two effects, described in Section II.2.1, on the leave-one-out estimator as on the resubstitution estimator. First consider the effect on variance: as the level of smoothing increases, the variance monotonically approaches zero. For any given type of smoothing and value of b, the smoothed U method has variance as large as or larger than the corresponding smoothed R method; however, the reduction in variance due to smoothing is generally greater for the U method than for the R method (see, for example, column 2 or column 5 of Table 3.3). Therefore, the effect of smoothing on the estimators' variances is actually more beneficial to the U method than to the R method.

The other effect of smoothing is that as the level of smoothing increases, the expected values of the estimators approach one-half. While this effect is often beneficial to the optimistically biased resubstitution estimator (see Section II.2.1), it is almost always detrimental to the leave-one-out estimator. Because the bias of the U method relative to the expected actual error rate is very small (Section I.2.3), changes in its expected value are very likely to increase the magnitude of the bias of this estimator.

In terms of UMSE, the U method is always improved by low levels of smoothing, indicating that the reduction in variance overshadows the increase in bias (see the third or sixth column of Table 3.3). As the level of smoothing increases, however, the UMSE also begins to increase.

This pattern is identical to that seen with the resubstitution estimator; however, there are important distinctions. First, the point at which the UMSE begins to rise is different for the U and R methods; sometimes it is greater for the U method than for the R method, and sometimes vice versa. Also, the minimum UMSE achieved by the smoothed resubstitution estimators always appears to be less than or equal to the minimum achieved by the smoothed leave-one-out estimators; this fact is not influenced by the relative values of UMSE for the unsmoothed U and R methods. For example, in Table 3.3, the minimum UMSE for the smoothed U method is approximately .014, while for the smoothed R method it is approximately .012.

Application of the idea of smoothing to the R method has two advantages over the application of it to the U method. While the bias of the R method with respect to the expected actual error rate usually tends to decrease with smoothing (at least at low levels), the bias of the U method tends to increase; and smoothed resubstitution estimators are potentially better than smoothed leave-one-out estimators in terms of UMSE. For these reasons, smoothed versions of the leave-one-out estimator are not considered any further.

III.2.2 The assumption of known variance

The assumption of known variance is usually artificial; being employed for theoretical convenience. This was the case with the numerical integration program. In most real situations, however, the variance of the parent distributions is not known. Since unknown variance is the situation of greatest interest, the results of the numerical integration study may be of limited value unless some measure of the effect of the assumption of known variance on the error rate estimators is obtained.

The numerical integration program is restricted to univariate normal parent distributions. In this case, if $c=0$, the resubstitution and leave-one-out estimators depend only on the sample means and not on s^2; therefore, they are identical with and without the assumption of known variance. To evaluate the effect of the assumption of known variance on the other estimators, and on the R and U methods for $c \neq 0$, the results of the Monte Carlo sampling program were reviewed. For example, in Table 3.3, the change in the UMSE of the estimators is very small (.001 or less). The mean and variance of the smoothed estimators are also not greatly affected by the assumption of known variance, while those of the parametric estimators are affected to a greater extent.

The general conclusion that the UMSEs of the error rate estimators are not greatly affected by the assumption of known variance is supported by all of the Monte Carlo sampling results. While the UMSEs of the parametric estimators decrease to a greater extent than those of the nonparametric estimators when known variance is assumed, these changes all seem to be relatively unimportant. Therefore, it is felt that the ordinality of the results of the numerical integration study are preserved in the case of unknown variance.

It is interesting to note that in the multivariate normal case the assumption of known variance has an entirely different effect. Consider, for example, Table 3.4, which corresponds to a five-variate normal situation. There, the decrease in the UMSE of the R method when assuming known variance is very large (.014), especially when compared to the decrease in the UMSE of the U method (.002); and the decrease for the D method is even greater (.017). Therefore, while the artificial assumption of known variance seems to have a benign effect in the univariate normal case, the same is not true in the multivariate normal case.

III.2.3 The effect of multivariate normality

In order to explore the effects of increasing the number of dimensions of the observation vector (r) on the UMSEs of the error rate estimators, the Monte Carlo program was run using three- and five-dimensional vectors, and with training samples of size ten and twenty-five. In this section these results are compared with the results from the univariate normal case.

TABLE 3.4

RESULTS OF THE MONTE CARLO SAMPLING PROGRAM

WITH 5-VARIATE NORMAL PARENT DISTRIBUTIONS,

$n_1=n_2=10$, $\Delta=1$, AND $c=0$

	Known Variance			Unknown Variance		
Estimator	Mean	Variance	UMSE	Mean	Variance	UMSE
R Method	.239	.012	.035	.210	.013	.049
U Method	.372	.022	.027	.389	.023	.029
D Method	.251	.004	.023	.216	.006	.040
DS Method	.291	.003	.015	.258	.005	.027
O Method	.256	.003	.022	.222	.006	.038
RS Method						
b=0.25	.243	.010	.031	.214	.011	.046
b=0.5	.250	.008	.028	.220	.009	.042
b=1.0	.275	.005	.020	.244	.007	.033
b=1.5	.307	.003	.013	.276	.005	.024
b=2.0	.339	.002	.010	.310	.004	.017
RN Method						
b=0.25	.248	.008	.029	.218	.010	.043
b=0.5	.266	.006	.022	.235	.007	.036
b=1.0	.312	.003	.013	.282	.005	.022
b=1.5	.351	.002	.009	.325	.003	.014
b=2.0	.379	.001	.009	.357	.002	.011
US Method						
b=0.25	.373	.019	.024	.389	.021	.027
b=0.5	.377	.017	.022	.391	.019	.025
b=1.0	.388	.012	.017	.397	.015	.021
b=1.5	.404	.008	.015	.407	.011	.018
b=2.0	.420	.005	.014	.418	.009	.017
UN Method						
b=0.25	.376	.017	.022	.390	.020	.026
b=0.5	.384	.013	.019	.395	.016	.022
b=1.0	.407	.008	.015	.409	.011	.018
b=1.5	.426	.005	.014	.424	.007	.016
b=2.0	.440	.003	.015	.436	.005	.015

While the UMSEs of all other error rate estimators always increase as r increases, the UMSE of the leave-one-out estimator can actually decrease. For example, compare Figures 3.1-3.3. When Δ is small, the U method has by far the largest UMSE In the univariate case, while in the multivariate cases, its UMSE is much improved relative to all other estimators. In fact, while the UMSEs for all other estimators at least double when r increases from one to five, the U method's UMSE decreases. These results indicate that, as the number of dimensions in the observation vector increases, the U method improves relative to all other estimators, expecially when Δ is small.

Increasing r also affects the UMSEs of the parametric estimators relative to one another. In the univariate case the UMSEs of the three parametric estimators are nearly equal (for example, Figure 3.1), while in the multivariate case they tend to separate (Figures 3.2 and 3.3). In almost all cases the DS method has the smallest UMSE, followed by the O method and then by the D method. The gaps between the parametric estimators seem to widen as the number of dimensions increases, but, for any given value of r, they seem to decrease as Δ increases.

Also important is how increasing dimensionality affects the smoothed estimators. In general, as r increases, larger values of b appear to be useful in reducing UMSE. To illustrate this effect, consider Figures 3.1-3.3, in which the five smooth slide estimators are displayed. In the univariate case (Figure 3.1), b=1.5 clearly results in too much smoothing (note the elevated UMSE for Δ>1.5), while the best results with respect to minimizing the UMSE seem to be obtained with b equal to one-half or one. In the three-dimensional case (Figure 3.2), b=1.5 appears to result in a good estimator,

FIGURE 3.1

UMSE EVALUATED BY MONTE CARLO SAMPLING

IN THE UNIVARIATE NORMAL CASE WITH

$n_1=n_2=25$, $c=0$, AND UNKNOWN VARIANCE

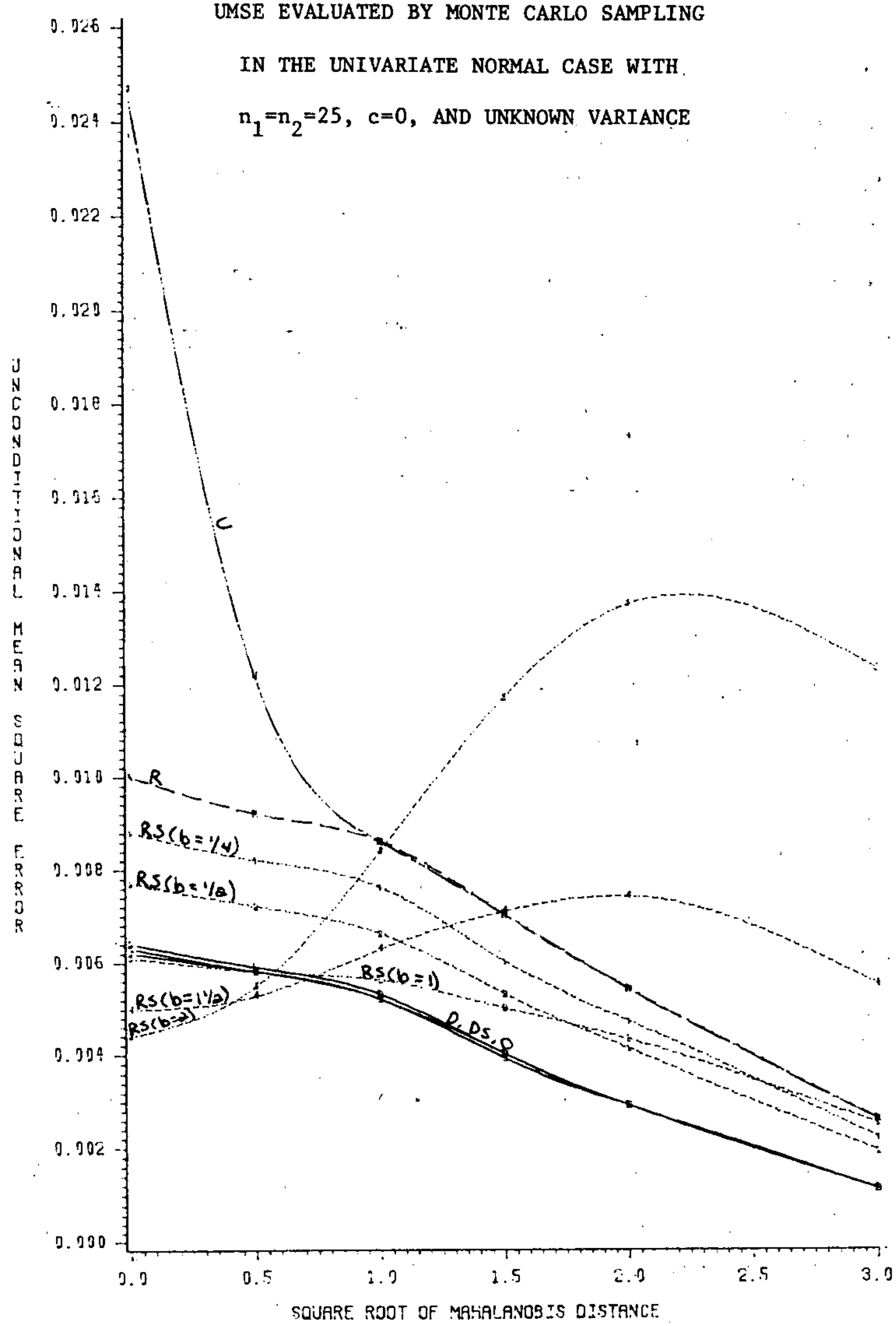


FIGURE 3.2

UMSE EVALUATED BY MONTE CARLO SAMPLING

IN THE 3-VARIATE NORMAL CASE WITH

$n_1=n_2=25$, c=0, AND UNKNOWN VARIANCE

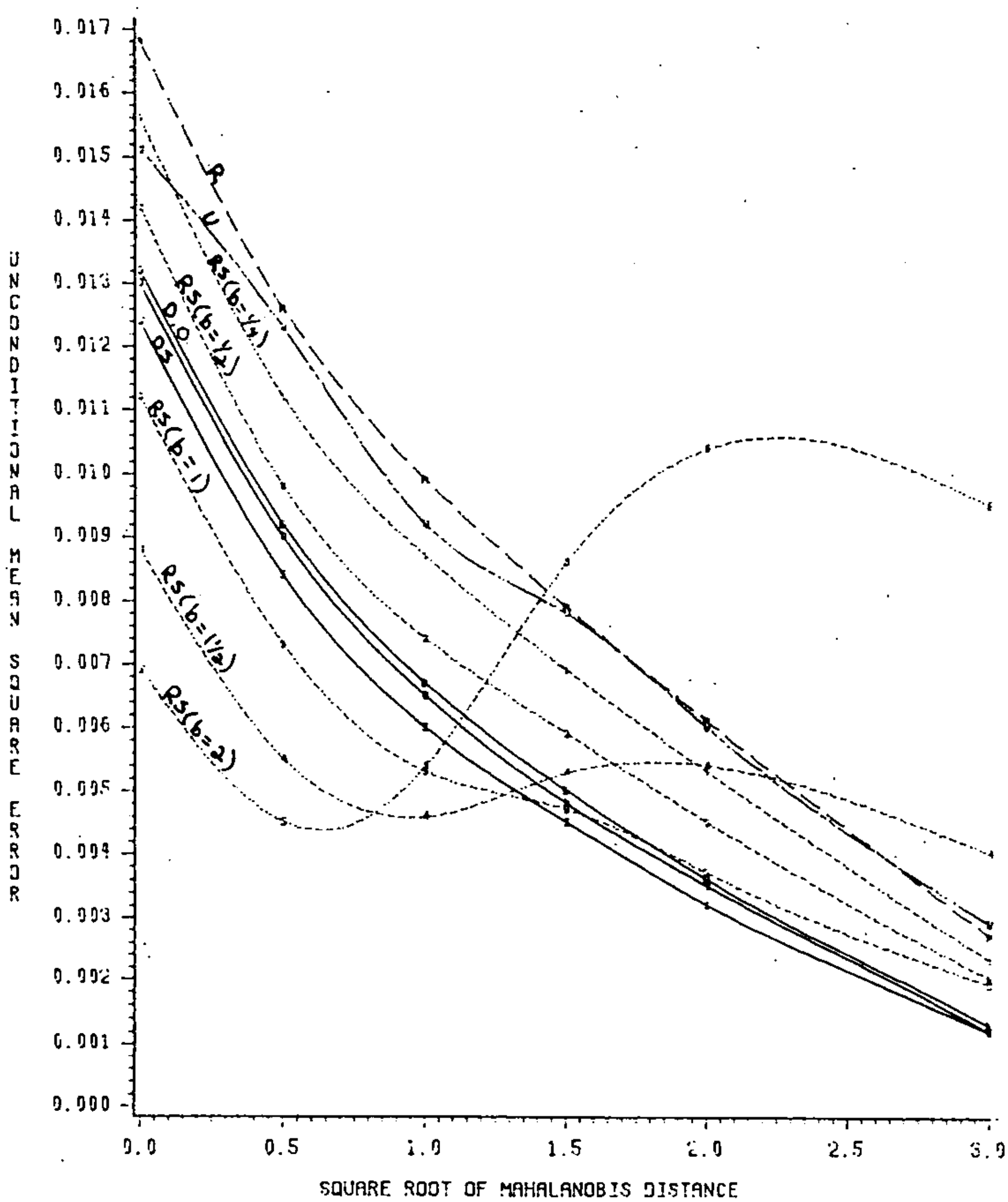
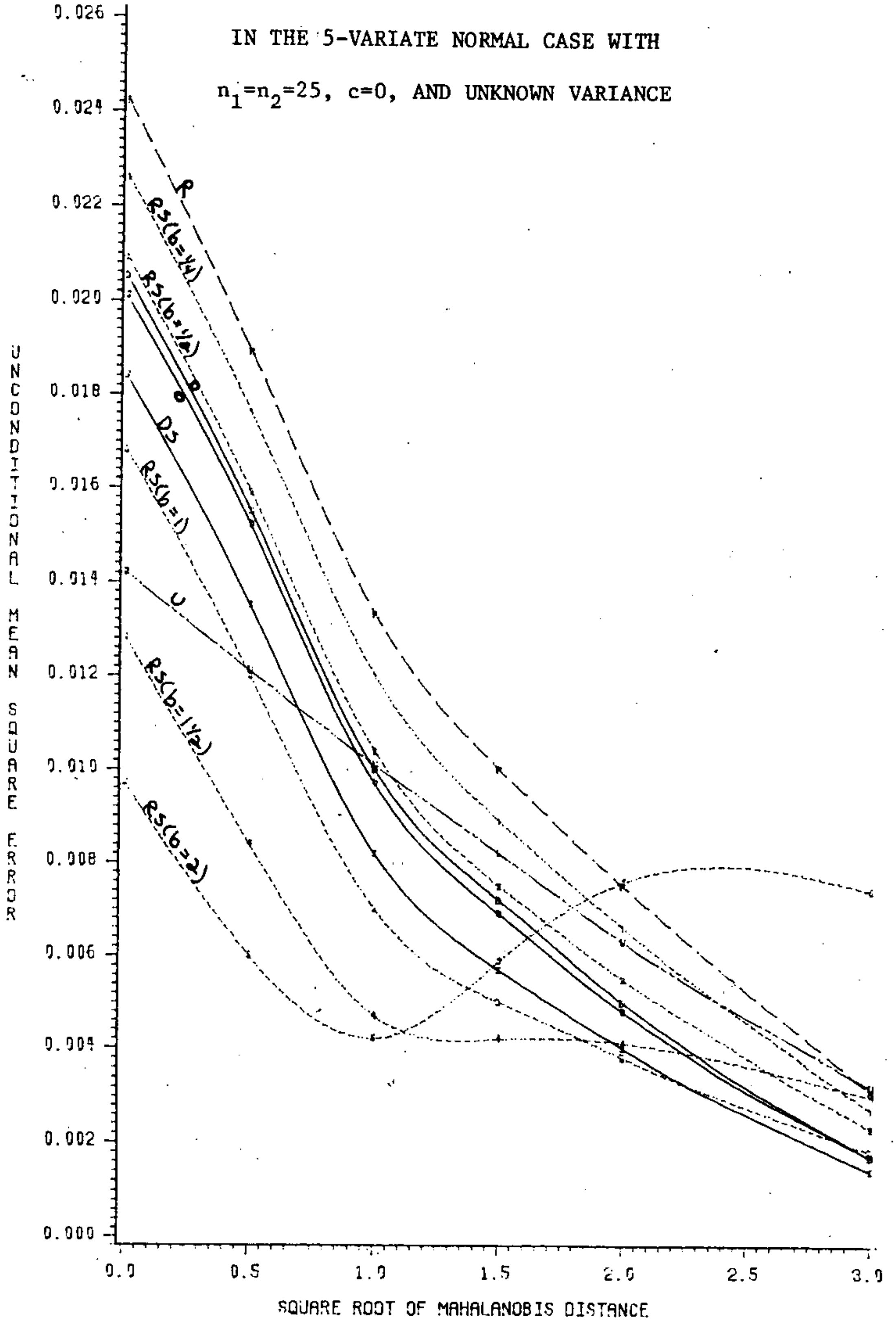


FIGURE 3.3

UMSE EVALUATED BY MONTE CARLO SAMPLING

IN THE 5-VARIATE NORMAL CASE WITH

$n_1=n_2=25$, $c=0$, AND UNKNOWN VARIANCE



although b=1 might still be preferred. In the five-dimensional case (Figure 3.3), however, b=1.5 gives excellent results. In every situation, at least one of the smoothed estimators compares very well with all other estimators with respect to minimizing UMSE. For example, in Figure 3.3, the smooth slide estimator with b equal to one and one-half has considerably lower UMSE than either the U or R methods over most of the range of Δ shown. Compared with the parametric estimators, this smoothed estimator has lower UMSE when Δ is less than two, and only slightly higher UMSE when Δ is greater than two.

III.2.4 The effect of non-normal parent distributions

The results presented in this chapter and in Chapter II indicate that the parametric estimators, the DS method in particular, tend to have lower UMSE than any of the nonparametric estimators. While in some instances one or more of the nonparametric estimators was shown to have smaller UMSE than the DS method for some values of Δ (for example, the U method in Figure 3.3), in no case was this true over the entire range of Δ. However, thus far only normally distributed parents have been considered; in this section the UMSEs of the error rate estimators are compared assuming three types of non-normal parents, described in Section I.5. The uniform and double exponential distributions were chosen as examples of symmetric distributions with tails narrower and wider, respectively, than those of the normal distribution; and the exponential distribution was chosen as an example of a skewed distribution. While it is expected that the U and R methods will prove to be more robust than the parametric estimators, it is of particular interest to determine whether or not the same is true of the smoothed estimators.

The results (some of which are illustrated in Figures 3.4-3.6) demonstrate that the UMSEs of some of the error rate estimators are relatively unaffected by non-normality, while those of the other estimators can be greatly affected. The estimators in the former class include the U and R methods and the smoothed estimators with low levels of smoothing (b equal to one-fourth or one-half in Figures 3.4-3.6). Compare these estimators in the normal case (Figure 3.1) with the same estimators in the three non-normal situations. The UMSEs of these estimators are higher in the uniform case than in the normal case, and lower in the exponential and double exponential cases. However, the differences between the normal and non-normal cases are always very small, and the UMSEs of these estimators relative to one another remain nearly identical.

The other group of estimators, including all of the parametric estimators and the smoothed estimators with high levels of smoothing (b equal to one and one-half or two in Figures 3.4-3.6), have UMSEs which are highly dependent on the class of parent distribution. In general, the UMSEs of the parametric estimators are larger under non-normality than under normality. The difference is especially large for those non-normal distributions with tails wider than those of the normal distribution. For example, the UMSEs for the parametric estimators are much larger in the exponential case in Figure 3.5 than in Figure 3.1. When the parent distributions are uniform (Figure 3.4), the difference is not very large; in fact, for small values of Δ, the UMSEs of the parametric estimators are smaller for uniform data than for normal data.

FIGURE 3.4

UMSE EVALUATED BY MONTE CARLO SAMPLING

IN THE UNIVARIATE UNIFORM CASE WITH

$n_1=n_2=25$, c=0, AND UNKNOWN VARIANCE

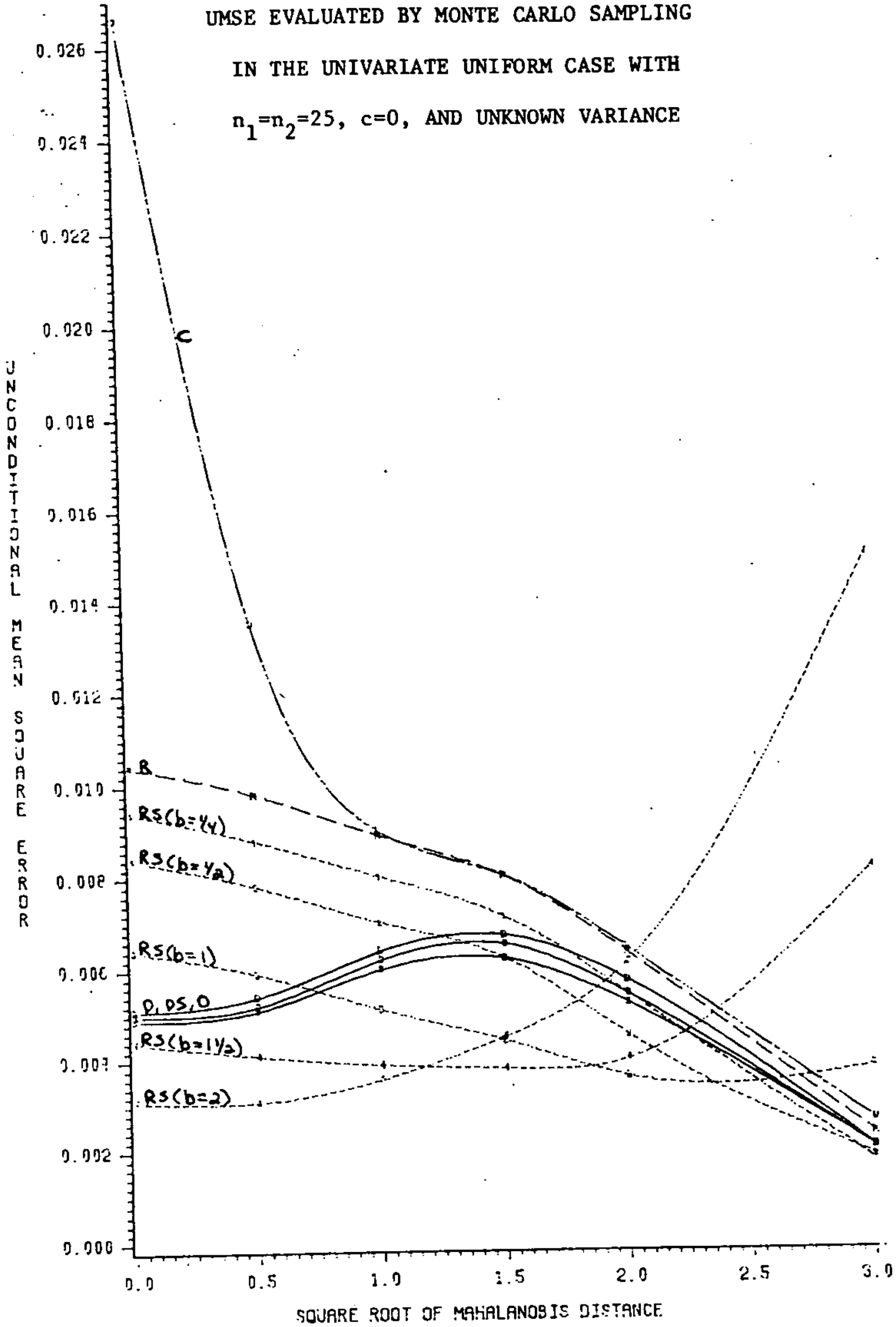


FIGURE 3.5

UMSE EVALUATED BY MONTE CARLO SAMPLING

IN THE UNIVARIATE EXPONENTIAL CASE WITH

$n_1=n_2=25$, c=0, AND UNKNOWN VARIANCE

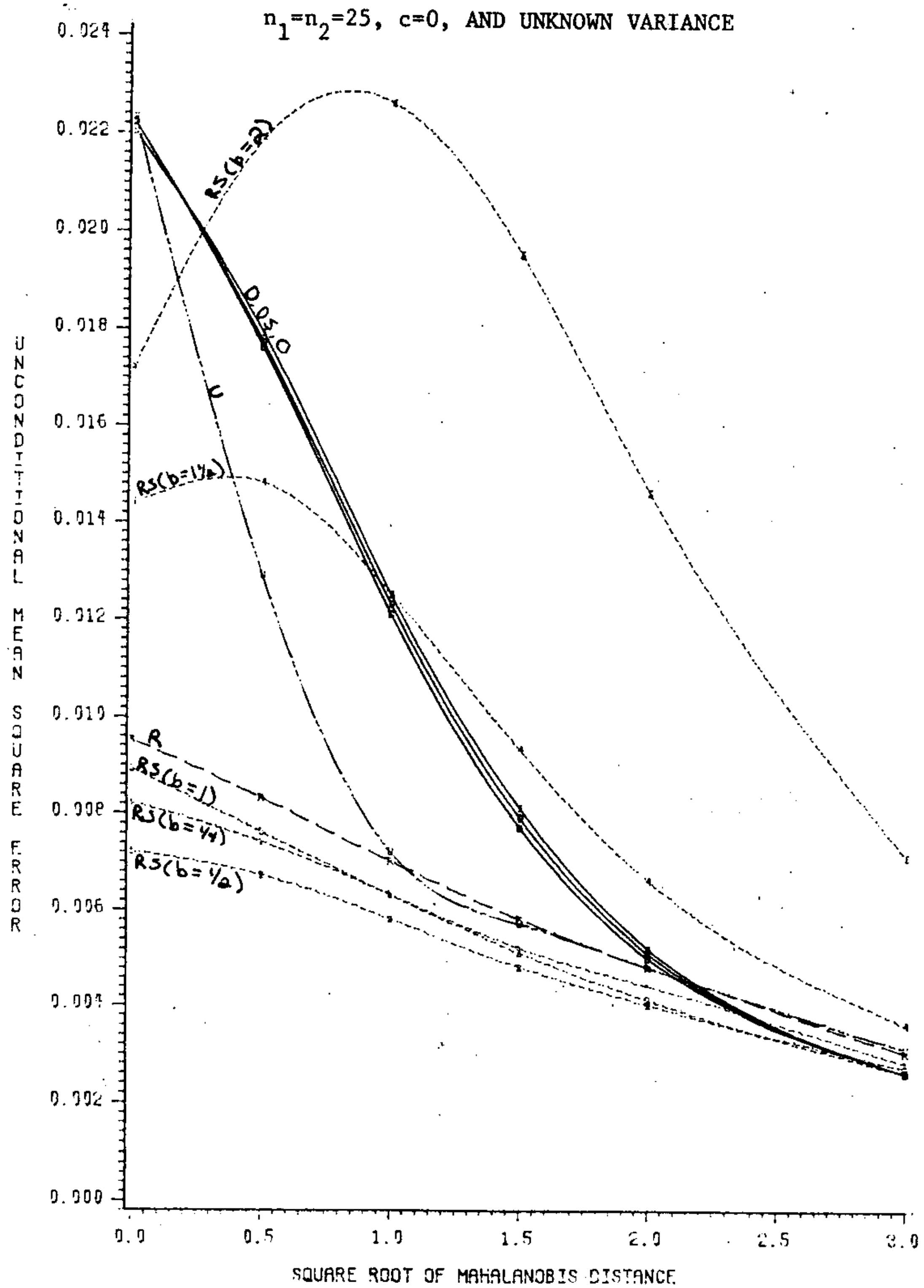
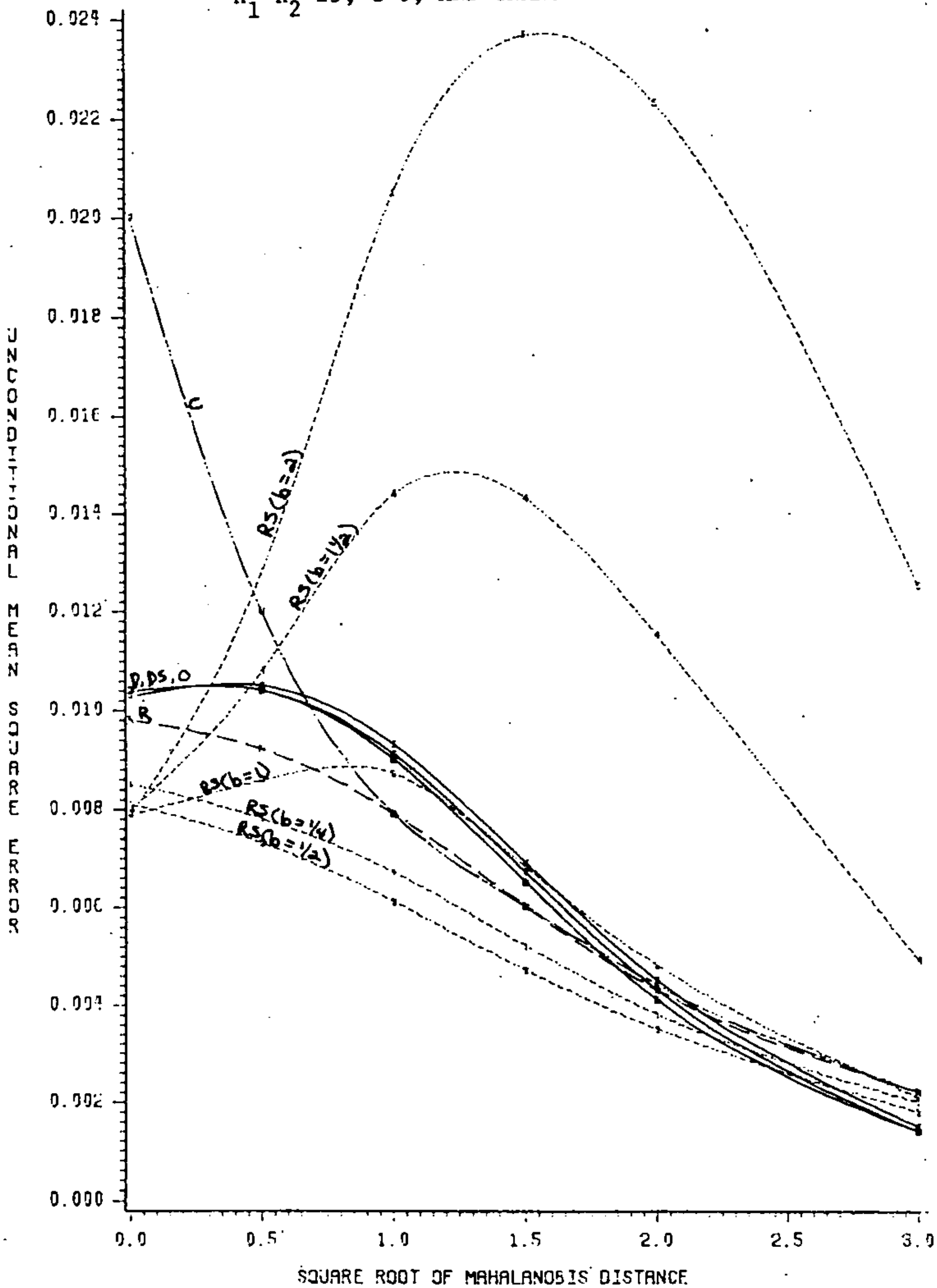


FIGURE 3.6

UMSE EVALUATED BY MONTE CARLO SAMPLING

IN THE UNIVARIATE DOUBLE EXPONENTIAL CASE WITH

$n_1=n_2=25$, c=0, AND UNKNOWN VARIANCE



The UMSEs of the smoothed estimators with large levels of smoothing seem to be affected by the class of parent distribution to an even larger extent than the parametric estimators. For example, the UMSEs of these estimators with uniform parent distributions in Figure 3.4 are entirely different than with exponential parent distributions in Figure 3.5. While the UMSEs of these smoothed estimators are sometimes smaller under non-normality than under normality (for example, uniform parent distributions with small values of Δ), they are usually larger, and are sometimes considerably larger.

While a low level of smoothing does not seem to affect the robustness of the smoothed estimator to a very large extent, a high level of smoothing certainly may. Because robustness is possibly the most important advantage of nonparametric estimators, the results of this section should serve as a caution against the use of too much smoothing. Therefore, any smoothed estimator suggested by this paper will be carefully evaluated in the three non-normal situations considered here.

III.2.5 Summary

One important result described in Section III.2 is that the assumption of known variance has very little effect on the UMSEs of the error rate estimators in the univariate normal case. Therefore, the results of the numerical integration study, described in Section II.2, are assumed to be valid with unknown variance as well. Two results of the multivariate normal case are that as r increases, 1) the DS method emerges as the parametric estimator with the smallest UMSE, followed by the O method; and 2) the UMSE of the U method improves dramatically relative to all other estimators', especially when Δ is small.

Replacing normal with non-normal parents has little effect on the UMSEs of the U and R methods and the smoothed estimators with low levels of smoothing; but may have a large effect on the parametric estimators and the smoothed estimators with high levels of smoothing.

The results of this section help to understand the properties of the smoothed error rate estimators. Smoothed leave-one-out estimators were found to be of less interest than the smoothed resubstitution estimators and are no longer being considered, but the smoothed resubstitution estimators were shown to have great potential. With the proper amount of smoothing these estimators were shown to have UMSE nearly as low as the parametric estimators in the univariate and multivariate normal cases. Under some types of non-normality, however, they have UMSE smaller than any other estimator.

It is clear that the smoothed estimators offer a reasonable alternative to the existing error rate estimators. The important question which remains to be answered is: How do we determine the proper level of smoothing? This question will be addressed in the next chapter, where a specific smoothed estimator, called the NS method, is defined. This estimator is then compared with the R and U methods, and the parametric estimators.

IV. THE NS METHOD: A PROPOSED SMOOTHED ESTIMATOR

In Chapters II and III the results of the numerical integration and Monte Carlo sampling studies were examined. The idea of smoothing in general was found to offer the possibility of a considerable reduction in the resubstitution estimator's UMSE. It was also found that the level of smoothing which maximizes this reduction in UMSE is a function of both the training sample sizes and the number of dimensions in the observation vector (r). In this chapter, a specific smoothed estimator is defined which it is felt achieves a large portion of the maximum reduction in UMSE. In Section IV.1 the development of this estimator, called the NS method, is detailed. The UMSE of the NS method is then compared with the UMSEs of the other estimators; the results are presented in Section IV.2.

IV.1 Definition of the NS Method

IV.1.1 Determination of the proper level of smoothing

Assume that an expression for the expected value of some type of smoothed estimator were known, conditioned on $\bar{\underset{\sim}{x}}_1$, $\bar{\underset{\sim}{x}}_2$, and $\underset{\sim}{S}$. This quantity, denoted $E(\hat{\alpha}^S \mid \bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2, \underset{\sim}{S})$, is a function of b, the smoothing constant. In Section II.2.1 it was shown that a low level of smoothing tends to reduce the bias of the resubstitution estimator as it simultaneously reduces the variance. Therefore, one way to obtain a reasonable level of smoothing would be to equate $E(\hat{\alpha}^S \mid \bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2, \underset{\sim}{S})$ with the expected actual error rate and solve for b. Unfortunately, this error rate is a function of the unknown parameters of the parent distributions; hence, this method is not feasible.

Another possibility is to equate $E(\hat{\alpha}^S \mid \bar{x}_1, \bar{x}_2, S)$ with the expected value of some other error rate estimator, also conditioned on $\bar{x}_1$, $\bar{x}_2$, and S. This estimator must both have an expected value which is independent of the unknown parameters of the parent distributions, and have less bias than the R method. Two such estimators are the D and DS methods.

It has been shown that the D method is less biased than the R method (Section I.2.2) and that the DS method tends to have less bias than the D method (Section I.2.5). Conditioned on $\bar{x}_1$, $\bar{x}_2$, and S, the D and DS methods are constant; hence their expected values under normality are

$$E(\hat{\alpha}^D \mid \bar{x}_1, \bar{x}_2, S) = \Phi\{-m/2\} \tag{4.1}$$

and

$$E(\hat{\alpha}^{DS} \mid \bar{x}_1, \bar{x}_2, S) = \Phi\{ -(n_1+n_2-r-3)^{\frac{1}{2}} m \,/\, 2(n_1+n_2-2)^{\frac{1}{2}} \}, \tag{4.2}$$

where $m^2 = (\bar{x}_1 - \bar{x}_2)' \, S^{-1} \, (\bar{x}_1 - \bar{x}_2)$.

The method used to determine a possible level of smoothing, therefore, is to determine an expression for $E(\hat{\alpha}^S \mid \bar{x}_1, \bar{x}_2, S)$, equate it with (4.1) or (4.2), and solve for b.

IV.1.2 An approximation for $E(\hat{\alpha}^S \mid \bar{x}_1, \bar{x}_2, S)$

According to Section II.2.2 it makes little difference which type of smoothing is used here in terms of UMSE. For mathematical convenience, normal smoothing is chosen rather than the smooth slide.

In Section II.1.5 an expression for $E(\hat{\alpha}^{RN} \mid \bar{x}_1, \bar{x}_2, S)$ was found in the univariate normal case with σ^2 assumed known. Throughout this section, however, it is assumed that the parent distributions are multivariate normal with means μ_1 and μ_2, and a common unknown variance, Σ. It is also assumed that the prior probabilities are equal, and hence c=0. Because the smoothed estimators are simply the average of a function of the independent sample observations, the following is true:

$$E(\hat{\alpha}^{RN} \mid \bar{x}_1, \bar{x}_2, S) = E[h^{RN}(X)] = E\{ \Phi[-w(X)/bm] \},$$

where $w(X) = [X - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)]' S^{-1} (\bar{x}_1 - \bar{x}_2)$ and X is a random observation from the training sample from π_1.

It can easily be shown that the joint distribution of X, $\bar{X}_1$, and $\bar{X}_2$ is

$$N \left[\begin{pmatrix} \mu_1 \\ \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma & n_1^{-1}\Sigma & 0 \\ n_1^{-1}\Sigma & n_1^{-1}\Sigma & 0 \\ 0 & 0 & n_2^{-1}\Sigma \end{pmatrix} \right],$$

and therefore that the distribution of X conditioned on $\bar{X}_1 = \bar{x}_1$ and $\bar{X}_2 = \bar{x}_2$ is

$$X \mid \bar{x}_1, \bar{x}_2 \sim N[\bar{x}_1 , (n_1 - 1)n_1^{-1}\Sigma].$$

Conditioned on $\bar{X}_1 = \bar{x}_1$, $\bar{X}_2 = \bar{x}_2$, and $\hat{\Sigma} = S$, X has expected value $\bar{x}_1$ and variance $(n_1-1)n_1^{-1}S$, but it does not have a normal distribution. It should be clear, however, that X in this case is asymptotically normal. Therefore, as an approximation, it will be assumed that

$$\underset{\sim}{X} \mid \bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2, \underset{\sim}{S} \sim N[\bar{\underset{\sim}{x}}_1 \;,\; (n_1-1)n_1^{-1}\underset{\sim}{S}]. \tag{4.3}$$

Now define $\underset{\sim}{Y} = [n_1/(n_1-1)]^{\frac{1}{2}}\, \underset{\sim}{S}^{-\frac{1}{2}}\, (\underset{\sim}{X}-\bar{\underset{\sim}{x}}_1)$, where $\underset{\sim}{S}^{\frac{1}{2}}$ is a square root decomposition of $\underset{\sim}{S}$. This gives $\underset{\sim}{X} = [(n_1-1)/n_1]^{\frac{1}{2}}\, \underset{\sim}{S}^{\frac{1}{2}}\, \underset{\sim}{Y} + \bar{\underset{\sim}{x}}_1$. If $\underset{\sim}{X}$ has the distribution given by (4.3), then $\underset{\sim}{Y} \sim N(\underset{\sim}{0} \;,\; \underset{\sim}{I}_r)$.

$$E(\hat{\alpha}^{RN} \mid \bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2, \underset{\sim}{S}) = E\{\, \Phi[-w(\underset{\sim}{X})/bm] \,\}$$

$$= E\left[\, \Phi\left\{ -\frac{[\underset{\sim}{X}-\frac{1}{2}(\bar{\underset{\sim}{x}}_1+\bar{\underset{\sim}{x}}_2)]'\; \underset{\sim}{S}^{-1}\; (\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2)}{bm} \right\} \right]$$

$$= E\left[\, \Phi\left\{ -\frac{[(n_1-1)^{\frac{1}{2}} n_1^{-\frac{1}{2}} \underset{\sim}{S}^{\frac{1}{2}} \underset{\sim}{Y} + \bar{\underset{\sim}{x}}_1 - \frac{1}{2}(\bar{\underset{\sim}{x}}_1+\bar{\underset{\sim}{x}}_2)]'\; \underset{\sim}{S}^{-1}\; (\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2)}{bm} \right\} \right]$$

$$= E\left[\, \Phi\left\{ -\frac{[(n_1-1)/n_1]^{\frac{1}{2}} \underset{\sim}{Y}' \underset{\sim}{S}^{-\frac{1}{2}} (\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2) + \frac{1}{2}m^2}{bm} \right\} \right]$$

$$= E[\, \Phi(T) \,], \text{ where } T \sim N[-m/2b \;,\; (n_1-1)/n_1 b^2]$$

$$= \Pr\{\, Z < T \,\}, \text{ where } Z \sim N(0,1)$$

$$= \Pr\{\, Z-T < 0 \,\}, \text{ where } Z-T \sim N[m/2b \;,\; 1 + (n_1-1)/n_1 b^2]$$

$$= \Phi\{\, -[1 + (n_1-1)/n_1 b^2]^{-\frac{1}{2}} m/2b \,\}$$

$$= \Phi\{\, -[n_1/(n_1 b^2 + n_1 - 1)]^{\frac{1}{2}} m/2 \,\}.$$

Because the distribution of $\underset{\sim}{X}$ used in this derivation is only approximate, the result is

$$E(\hat{\alpha}^{RN} \mid \bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2, \underset{\sim}{S}) \doteq \Phi\{\, -[n_1/(n_1 b^2 + n_1 - 1)]^{\frac{1}{2}} m/2 \,\}. \tag{4.4}$$

IV.1.3 Solution for b

In order to obtain a normal smoothing estimator with expected value equal to that of the D method, (4.1) is equated with (4.4). This yields

$$\Phi\{-m/2\} = \Phi\{\ -[n_1/(n_1 b^2+n_1-1)]^{\frac{1}{2}} m/2\ \}$$

or

$$1 = n_1/(n_1 b^2+n_1-1)$$

or

$$b = n_1^{-\frac{1}{2}}. \tag{4.5}$$

Similarly, in order to obtain a normal smoothing estimator with expected value approximately equal to that of the DS method, (4.2) is equated with (4.4). This yields

$$\Phi\{-[(n_1+n_2-r-3)/(n_1+n_2-2)]^{\frac{1}{2}} m/2\ \} = \Phi\{\ -[n_1/(n_1 b^2+n_1-1)]^{\frac{1}{2}} m/2\ \}$$

or

$$(n_1+n_2-r-3)/(n_1+n_2-2) = n_1/(n_1 b^2+n_1-1)$$

or

$$b = \{\ [(r+2)(n_1-1) + (n_2-1)]\ /\ [n_1(n_1+n_2-r-3)]\ \}^{\frac{1}{2}}. \tag{4.6}$$

The smoothing constant defined by (4.6) has certain advantages over that defined by (4.5). It was shown in Section III.2.3 that r is an important factor in the choice of the level of smoothing, but, while the constant defined by (4.6) is influenced by r, the constant defined by (4.5) is not. An examination of the results of the Monte Carlo sampling study reveals the following. The smoothing constant in (4.5) yields an estimator which always has lower UMSE than the R method; however, the reduction in UMSE due to smoothing is usually very small. This estimator does not seem to smooth enough. The smoothing constant in (4.6) is larger than that in (4.5); hence this estimator has a larger level of smoothing than the other smoothed estimator. The reduction in UMSE due to smoothing is usually quite large for this estimator.

The smoothed estimator proposed here, called the NS estimator, is defined as follows. For each observation in the sample from π_1, calculate

$$h^{NS}(\underset{\sim}{x}_i) = \Phi\{ -[n_1(n_1+n_2-r-3)]^{\frac{1}{2}} w(\underset{\sim}{x}_i) \ / \ m[(r+2)(n_1-1) + (n_2-1)]^{\frac{1}{2}} \},$$

and take the mean of the n_1 values.

IV.2 Evaluation of the NS Method

The results of the same numerical integration and Monte Carlo sampling runs which were discussed in Chapters II and III are used to compare the NS method with other error rate estimators. In addition, in order to obtain a wider range of situations under which the NS method could be evaluated, several new runs were performed. These were listed in Table 1.3.

The numerical integration and Monte Carlo programs evaluate the smoothed estimators only at five specific values of b; however, the value of b defined by (4.6) is usually not among them. In order to estimate the results for the NS method for these programs, a quadratic interpolation formula, based on Newton's divided difference interpolating polynomial (Carnahan, Luther and Wilks 1969, p. 9) is employed. This method is defined as follows. Let b_i, $i=1,\ldots,5$, correspond to the five levels of smoothing which are considered by the two programs, in ascending order. Therefore, $b_1=0.25$, $b_2=0.5$, and so on. Similarly, let $b_0=0$ correspond to no smoothing: the R method. Now let $f(b_i)$, $i=0,\ldots,5$, denote one of the three functions evaluated by the programs (mean, variance, or UMSE). In order to estimate f(b) for b defined as in (4.6), determine the integer j such that $1 \leq j \leq 4$ and $|b-b_j| \leq |b-b_i|$ for $1 \leq i \leq 4$; and calculate

$$f(b) \doteq f(b_{j-1}) + (b-b_{j-1})[f(b_j) - f(b_{j-1})] / (b_j-b_{j-1})$$
$$+ (b-b_{j-1})(b-b_j) \{ [f(b_{j+1}) - f(b_j)] / (b_{j+1}-b_j)$$
$$- [f(b_j) - f(b_{j-1})] / (b_j-b_{j-1}) \} / (b_{j+1}-b_{j-1}).$$

In every case considered, the value of b defined by (4.6) is within the range of smoothing constants considered by the programs; in other words, $0 < b < 2$. The quadratic interpolation formula is used rather than a simple linear interpolation for greater accuracy, especially in light of the fact that the functions being interpolated generally do not appear to be linear. However, the difference between the results obtained using linear and quadratic interpolation formulas is always very small. For example, consider the results for the RN method presented in Table 3.4. The sixth column of this table corresponds to the UMSE of the estimator when the variance of the parent distributions is unknown. Using the quadratic interpolation formula to this function for b=0.7746 yields 0.0276, while a linear interpolation yields 0.0283.

IV.2.1 Results in the univariate normal case

Figures 4.1-4.4 display the UMSEs of five error rate estimators in the univariate normal case with four different sets of training sample sizes. Of the estimators considered in Chapters II and III, the smoothed estimators are here represented by the NS method, and and the 0 method is omitted, since it was found to have UMSE nearly

FIGURE 4.1

UMSE EVALUATED BY NUMERICAL INTEGRATION

IN THE UNIVARIATE NORMAL CASE WITH

$n_1=n_2=5$, c=0, AND KNOWN VARIANCE

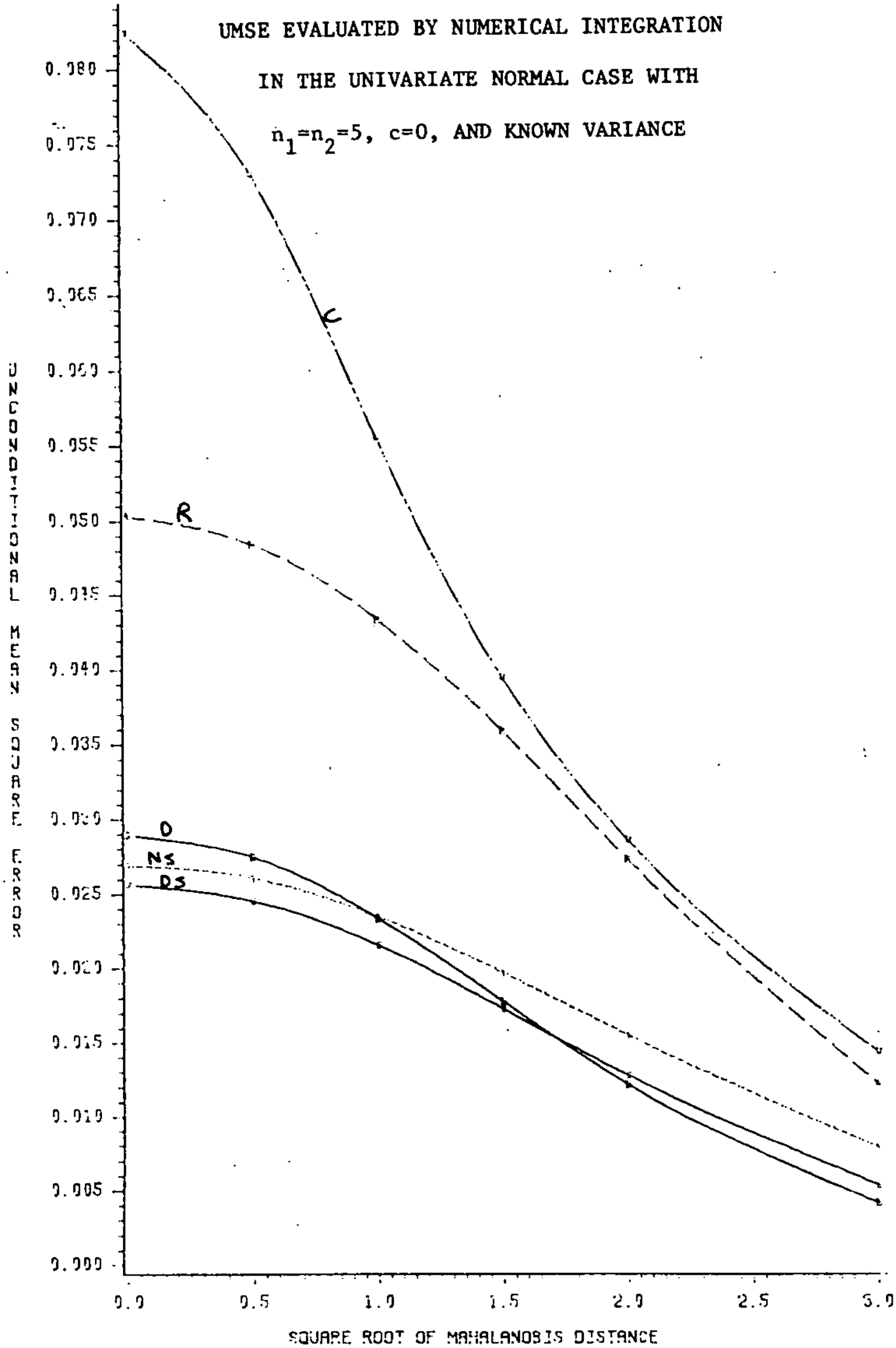


FIGURE 4.2

UMSE EVALUATED BY NUMERICAL INTEGRATION

IN THE UNIVARIATE NORMAL CASE WITH

$n_1=n_2=10$, c=0, AND KNOWN VARIANCE

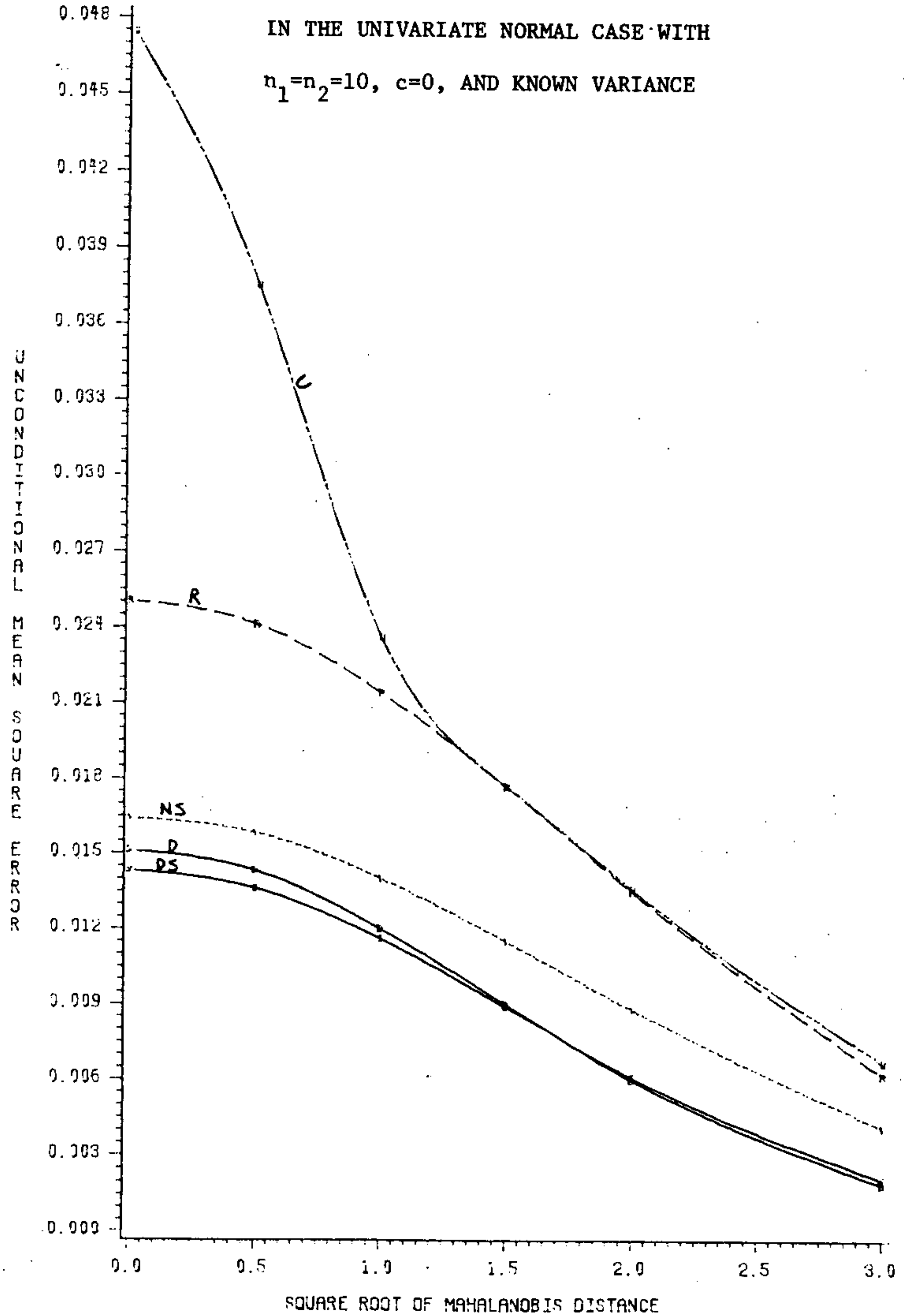


FIGURE 4.3

UMSE EVALUATED BY NUMERICAL INTEGRATION

IN THE UNIVARIATE NORMAL CASE WITH

$n_1=n_2=25$, c=0, AND KNOWN VARIANCE

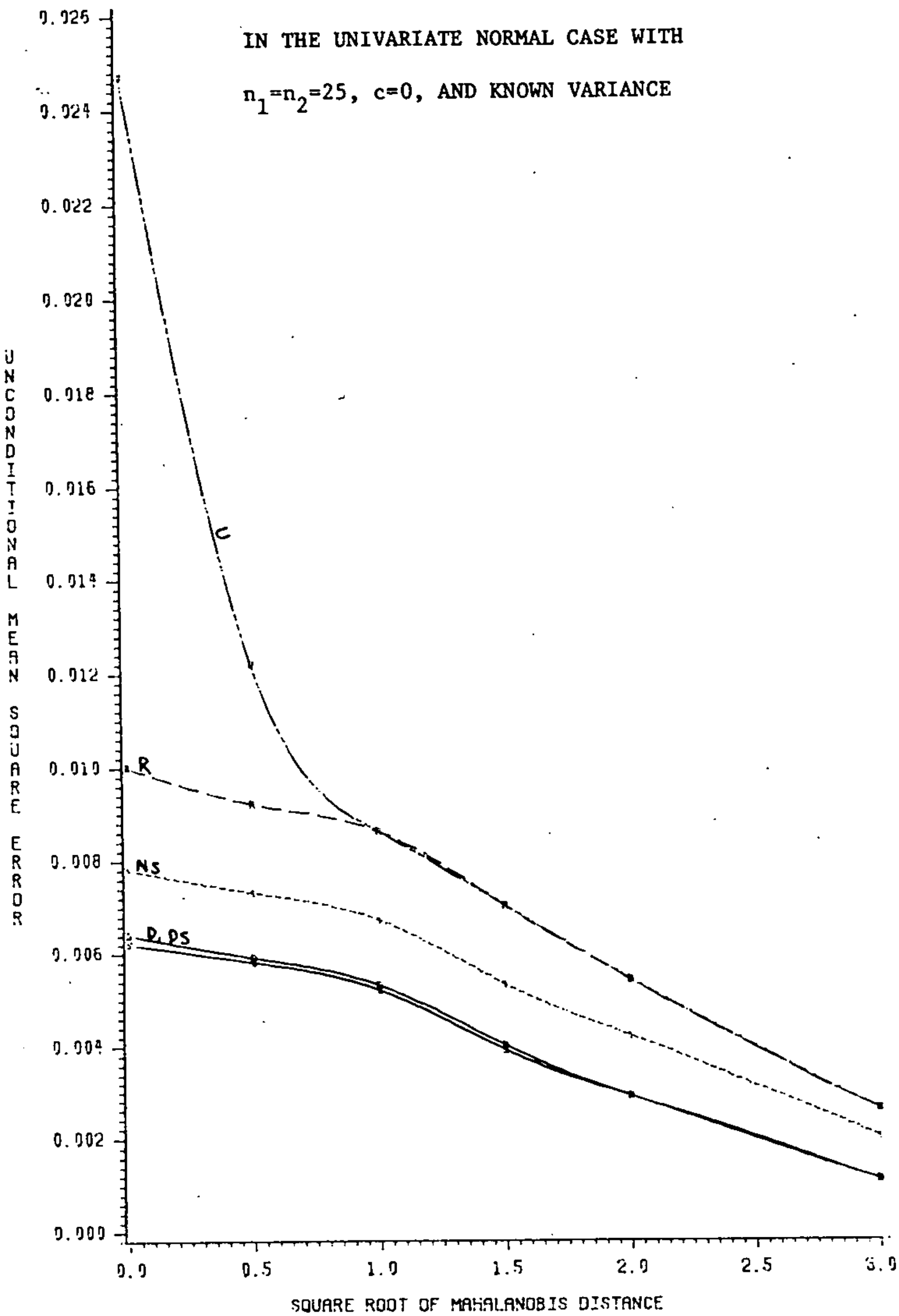
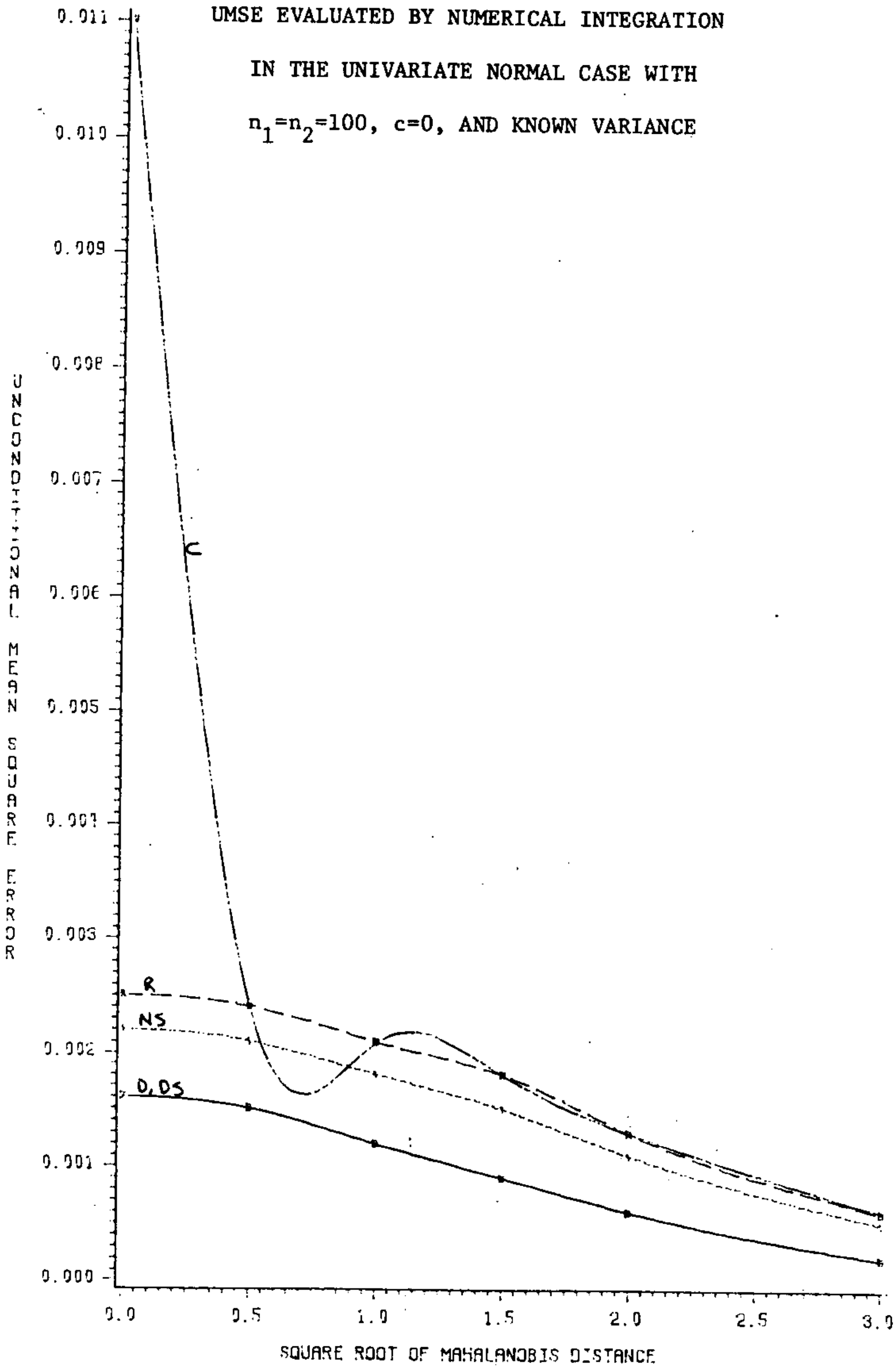


FIGURE 4.4

UMSE EVALUATED BY NUMERICAL INTEGRATION

IN THE UNIVARIATE NORMAL CASE WITH

$n_1=n_2=100$, c=0, AND KNOWN VARIANCE



always intermediate to those of the D and DS methods. The results in these figures were obtained from the numerical integration program in order to conveniently cover a very wide range of training sample sizes. In Section III.2.2 it was found that the assumption of known variance implicit in these results has only a very small effect.

It is interesting to consider how the relationships among the UMSEs of the estimators vary with the training sample sizes. As the training sample sizes increase, the UMSEs of all five estimators decrease, while the rank order of these UMSEs remains nearly constant. With one exception, the NS method has UMSE greater than both the D and DS methods', and smaller than both the U and R methods'. This exception occurs with training samples of size five (Figure 4.1), where, with Δ less than one, the UMSE of the NS method is less than that of the D method.

For the smallest training sample sizes, the UMSE of the NS method is much closer to those of the parametric estimators than it is to that of the R method. For the largest training sample sizes, however, the opposite is observed. Figures 4.1-4.4 illustrate a trend as the training sample sizes increase: the difference between the UMSEs of the NS and parametric methods increases relative to the difference between the UMSEs of the NS and R methods. This result indicates, therefore, that smoothing improves the R method most dramatically in just those situations where the UMSEs are largest, and, hence, where the improvement is most beneficial.

IV.2.2 Results in the multivariate normal case

When comparing the UMSE of the NS method with those of the other error rate estimators, the relationships observed in the multivariate normal case are very similar to those that were observed in the univariate normal case. For the smallest training sample sizes, the UMSEs of the NS and DS methods are nearly equal, and both are smaller than the UMSEs of the D and R methods. Again, this indicates that smoothing is most effective when the UMSEs are largest, and hence, where it is most necessary.

For the largest training sample sizes, illustrated by Figures 4.5 and 4.6, the NS method has UMSE which is less than the UMSE of the R method and greater than the UMSE of the DS method. In Figure 4.5, with r=3, the UMSE of the NS method is greater than that of the D method, while in Figure 4.6, with r=5, the opposite is observed.

IV.2.3 Results in non-normal situations

In Section III.2.4 it was learned that if the level of smoothing is small, the smoothed estimators, like the R and U methods, are robust; that is, their UMSEs are relatively unaffected by non-normal parent distributions. However, it was also found that the UMSEs of the smoothed estimators could be greatly affected by non-normality if the level of smoothing is large. Therefore, it is important to investigate the question of robustness as it relates to the NS method.

In general, the UMSE of the NS method does not seem to be greatly affected by non-normality. For example, consider the results for three non-normal distributions presented in Figures 4.7-4.9, along with the results under normality in Figure 4.3. For any given value of Δ, the UMSE of the NS does not seem to vary among these figures any more than the UMSEs of the R and U methods. Apparently, the level of smoothing in the NS method is not large enough to appreciably affect this estimator's robustness.

The UMSEs of the parametric estimators are, of course, greatly affected by the parent distribution type. The result is that in some situations (Figure 4.7) the parametric estimators have UMSE lower than that of the NS method, while in other situations (Figures 4.8 and 4.9) the NS method has UMSE considerably lower than that of any other estimator studied.

FIGURE 4.5

UMSE EVALUATED BY MONTE CARLO SAMPLING

IN THE 3-VARIATE NORMAL CASE WITH

$n_1=n_2=25$, c=0, AND UNKNOWN VARIANCE

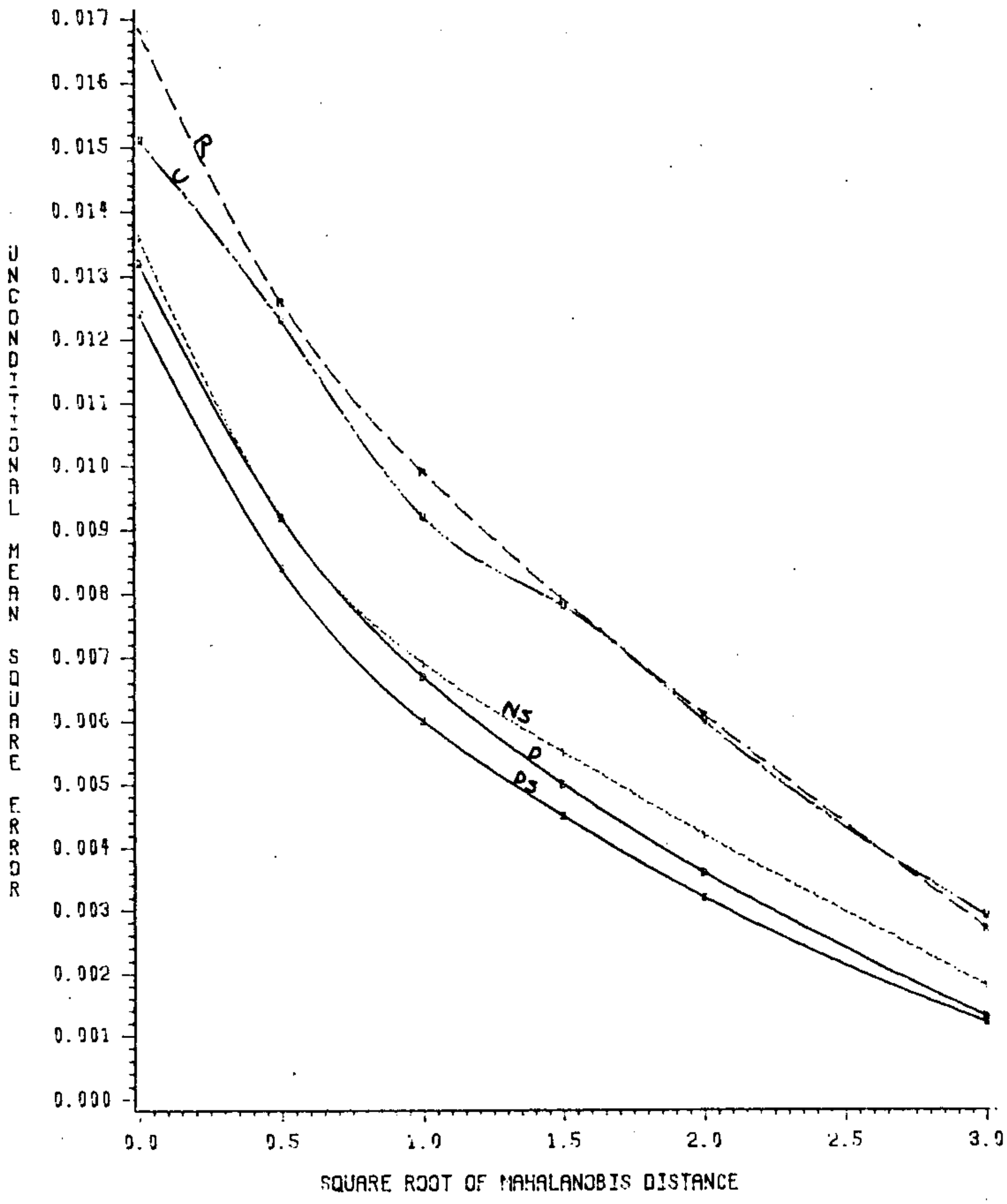


FIGURE 4.6

UMSE EVALUATED BY MONTE CARLO SAMPLING

IN THE 5-VARIATE NORMAL CASE WITH

$n_1=n_2=25$, c=0, AND UNKNOWN VARIANCE

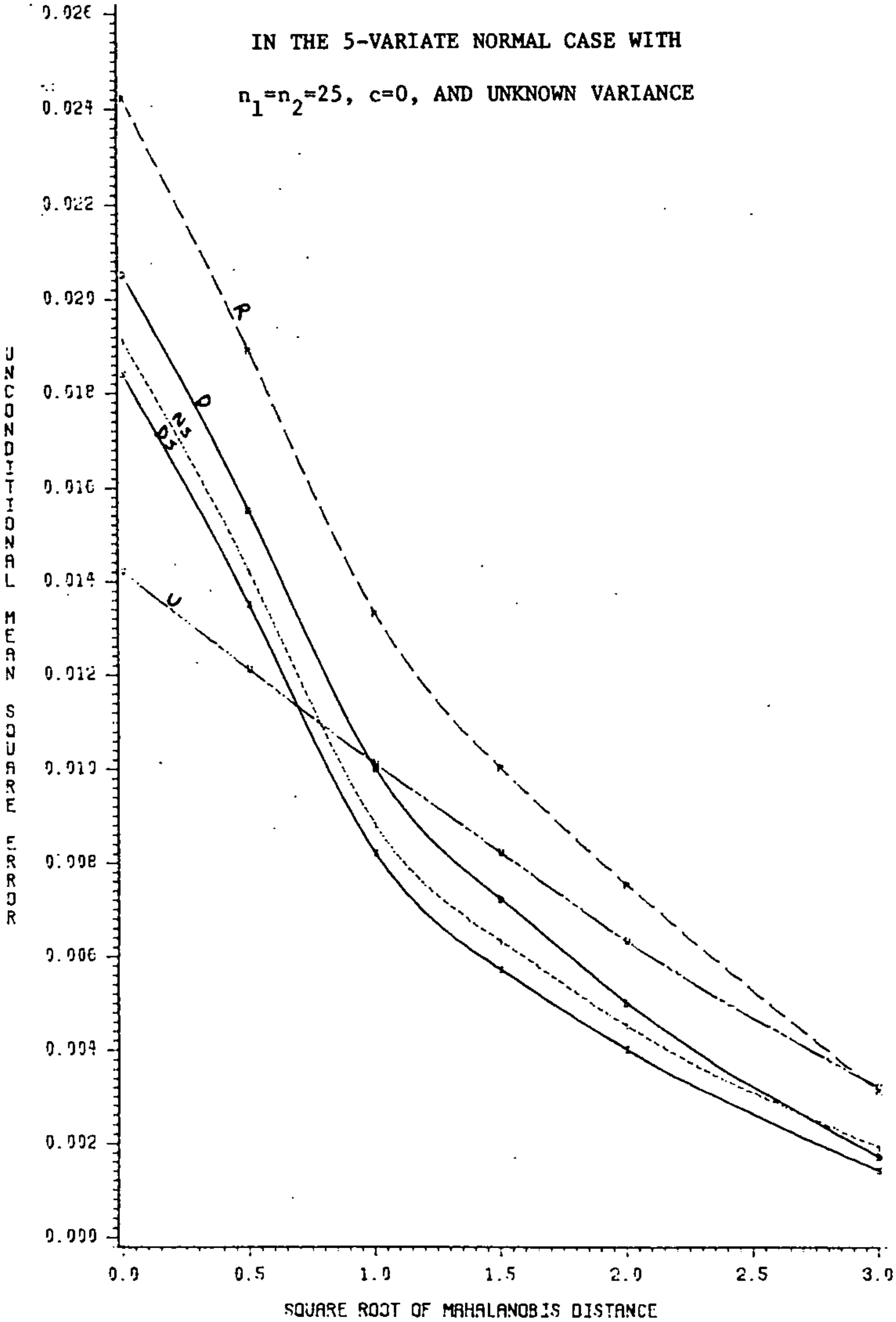


FIGURE 4.7

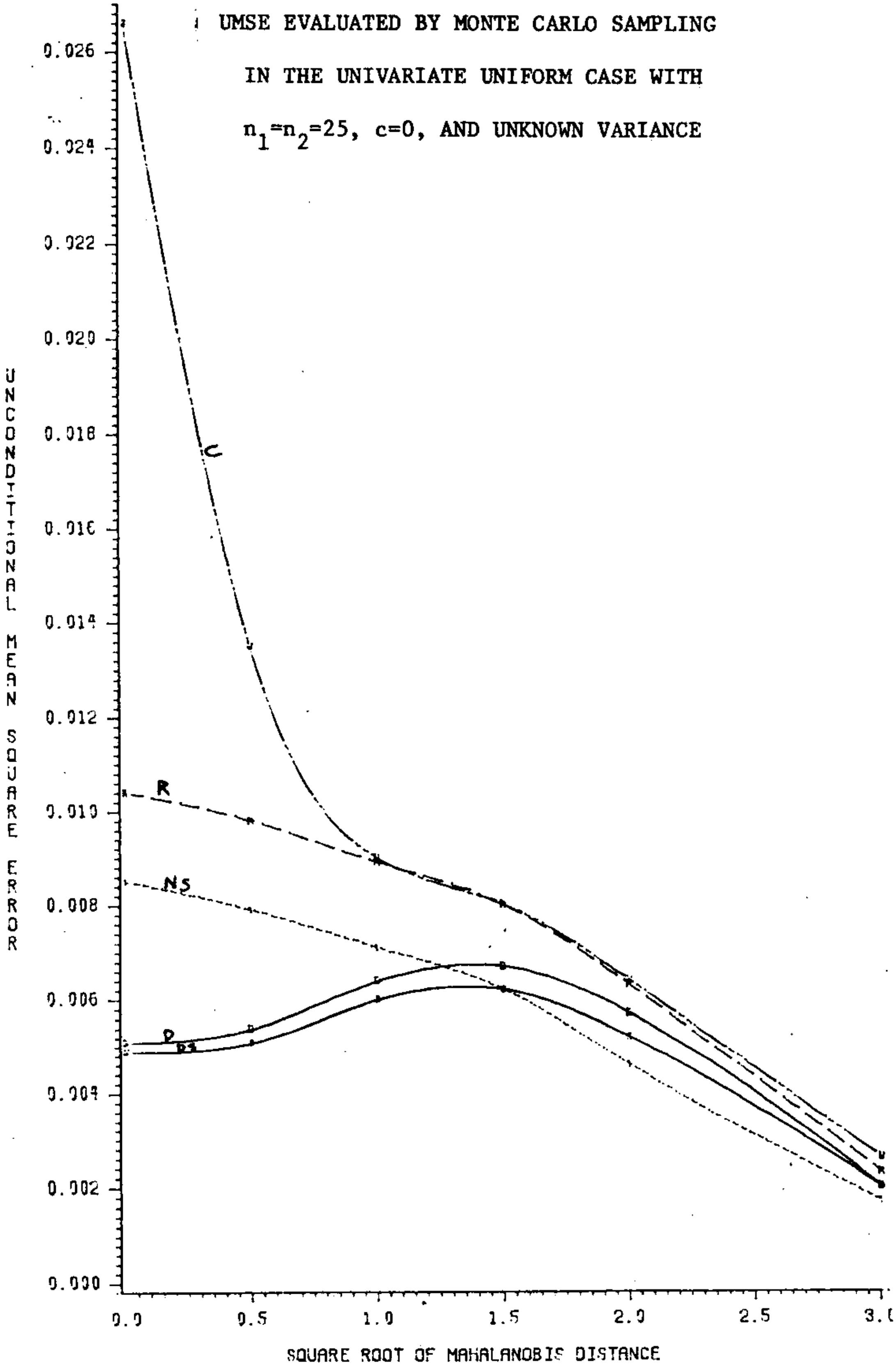

UMSE EVALUATED BY MONTE CARLO SAMPLING
IN THE UNIVARIATE UNIFORM CASE WITH
$n_1=n_2=25$, c=0, AND UNKNOWN VARIANCE
UNCONDITIONAL MEAN SQUARE ERROR
0.026
0.024
0.022
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
C
R
NS
D
0.0
0.5
1.0
1.5
2.0
2.5
SQUARE ROOT OF MAHALANOBIS DISTANCE

FIGURE 4.8

UMSE EVALUATED BY MONTE CARLO SAMPLING

IN THE UNIVARIATE EXPONENTIAL CASE WITH

$n_1=n_2=25$, c=0, AND UNKNOWN VARIANCE

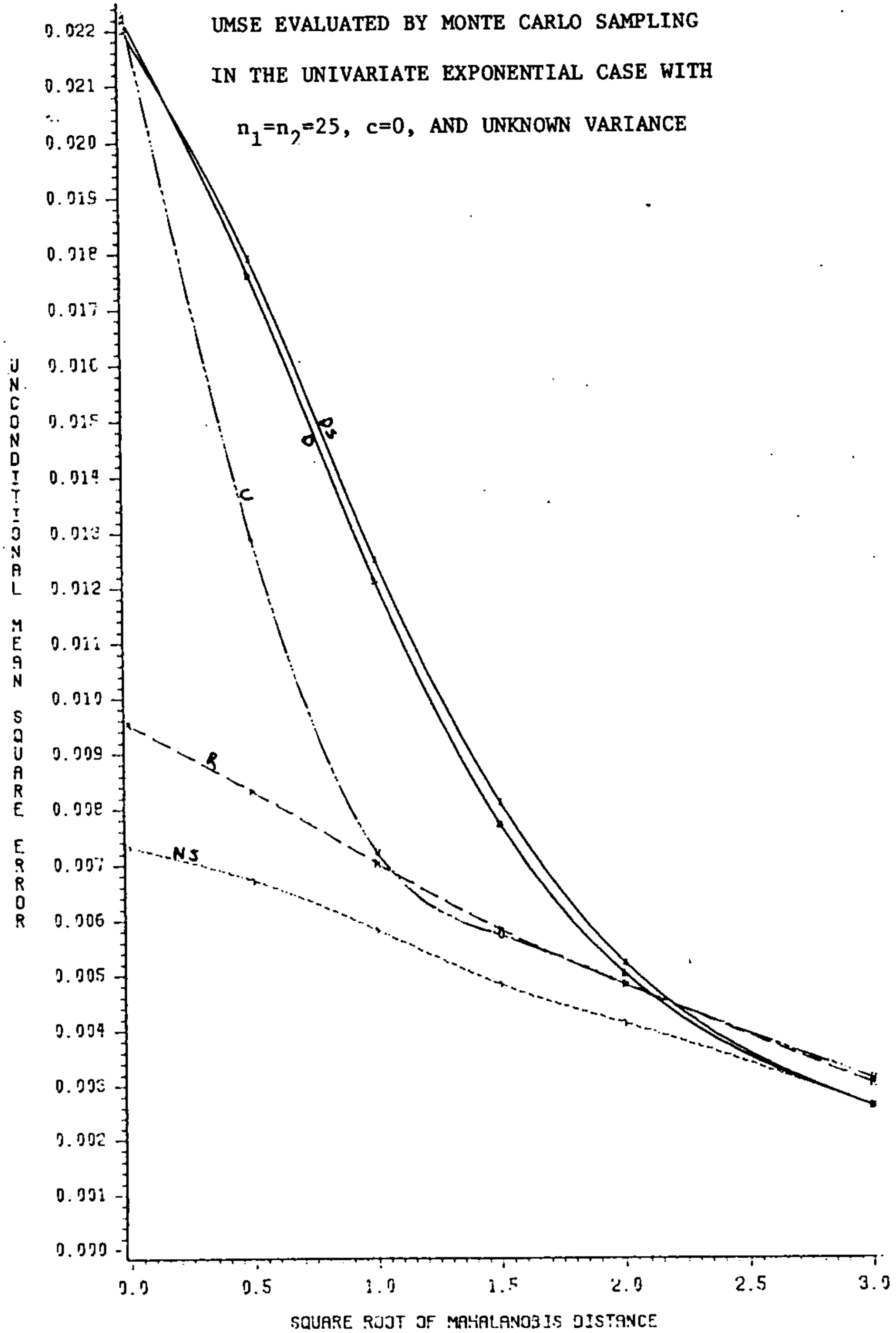
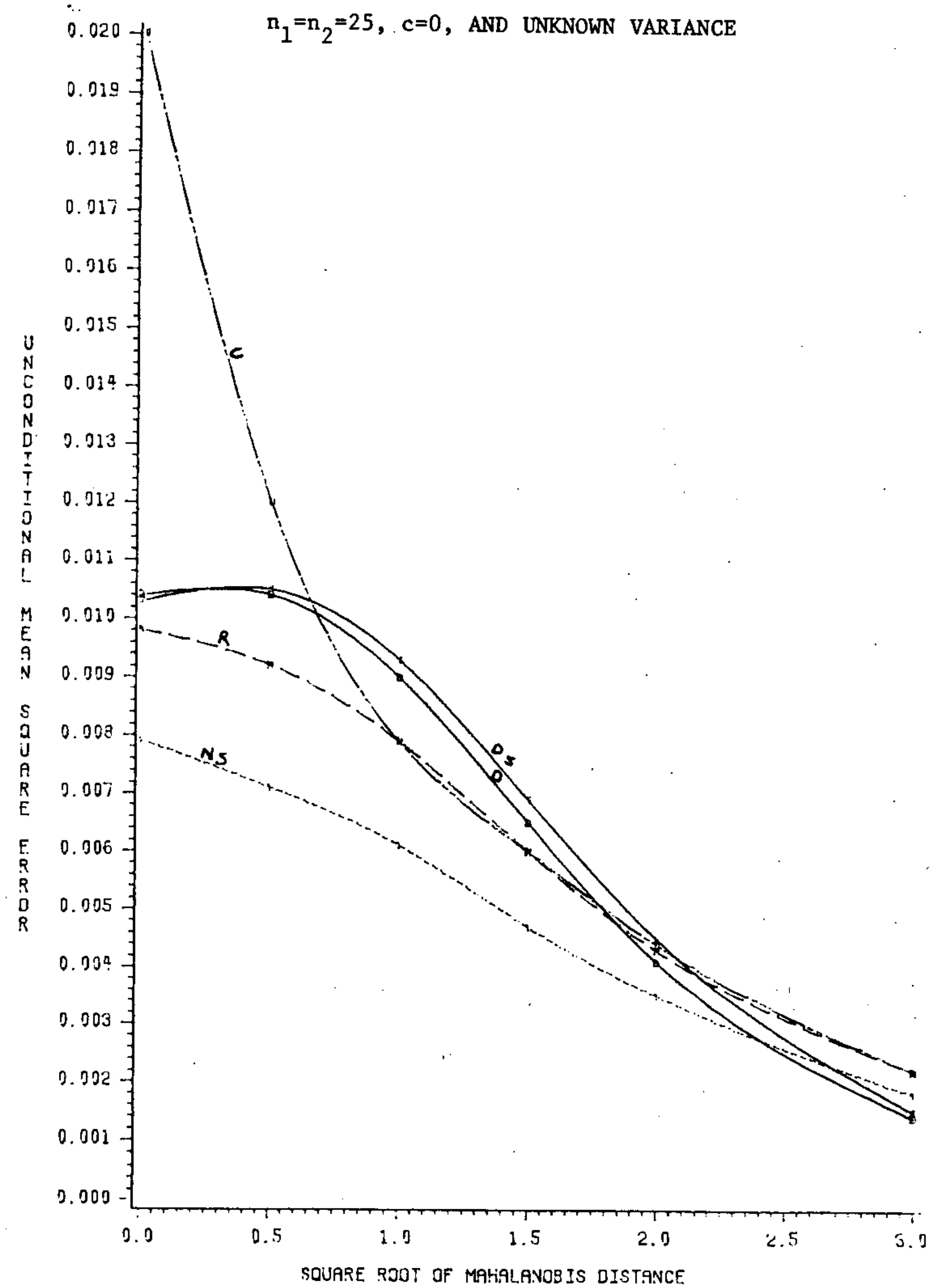


FIGURE 4.9

UMSE EVALUATED BY MONTE CARLO SAMPLING

IN THE UNIVARIATE DOUBLE EXPONENTIAL CASE WITH

$n_1=n_2=25$, c=0, AND UNKNOWN VARIANCE



V. AN APPLICATION OF THE NS METHOD

The NS method has been shown to perform very well relative to other error rate estimators in many situations when the linear discriminant function (LDF) is used. It is also important to demonstrate how this method may be applied when discrimination functions other than the LDF are used. In this chapter, therefore, the NS method is used to predict the error rates of several different discriminant rules developed from a real data set.

Luria et al. (1976), studying two-year survival following myocardial infarction, collected data on training samples of 110 patients who lived and 27 who died, and testing samples of 91 who lived and 14 who died. A stepwise discriminant procedure was performed on the training samples, and resulted in five variables being selected as minimally sufficient for prediction. These include two continuous variables (z_1 and z_2: systolic blood pressure and blood urea nitrogen) and three binary variables (y_1, y_2, and y_3: presence of atrial arrhythmias, history of coronary artery disease, and more than one ventricular ectopic beat per hour) scored as 1 for absent and 2 for present.

Knoke (1982) analyzed these data using four types of discriminant rules developed from the training samples: the LDF, logistic regression (LR), the quadratic discriminant function (QDF), and the location model (LM). Error rates for these rules were estimated based on the resubstitution and leave-one-out methods; these were compared with unbiased estimates of the error rates obtained from the independent testing samples. The purpose here is to reproduce some of this work, adding error rate estimates based on the NS method.

Logistic regression is frequently used in discrimination problems, especially when one or more of the explanatory variables is categorical. The application of the NS method is essentially identical when logistic regression and linear discrimination are used. To illustrate this fact, consider that logistic regression and linear discrimination are simply two distinct methods of estimating the same parameters.

In linear discrimination, multivariate normal observation vectors are assumed; hence, the log of the likelihood function can be written as

$$\ln L(\underset{\sim}{x}) = \alpha_0 + \underset{\sim}{x}'\underset{\sim}{\alpha},$$

where $\alpha_0 = -c - \frac{1}{2}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)' \underset{\sim}{\Sigma}^{-1} (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)$ is the intercept term, $\underset{\sim}{\alpha} = \underset{\sim}{\Sigma}^{-1} (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)$ is a vector of parameters, $c = \ln(q_2/q_1)$, and q_1 and q_2 are the prior probabilities. In the LDF these parameters are estimated by $a_0 = -\hat{c} - \frac{1}{2}(\bar{\underset{\sim}{x}}_1+\bar{\underset{\sim}{x}}_2)' \underset{\sim}{S}^{-1} (\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2)$, $\underset{\sim}{a} = \underset{\sim}{S}^{-1} (\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2)$, $\hat{c} = \ln(\hat{q}_2/\hat{q}_1)$, and $\hat{q}_1$ and $\hat{q}_2$ are estimates of the prior probabilities. In logistic regression, a different method of parameter estimation is used and different estimates are obtained, but the same parameters are estimated, and a linearized classification rule may be written.

There is one important distinction between the estimates of α_0 in linear discrimination and logistic regression. While in linear discrimination any desired estimates, $\hat{q}_1^*$ and $\hat{q}_2^*$, of the prior probabilities may be chosen, in logistic regression these values are automatically estimated by the relative frequencies of the training samples: hence, in logistic regression, $\hat{q}_1 = n_1/(n_1+n_2)$ and $\hat{q}_2 = n_2/(n_1+n_2)$. In order to use estimates of the prior probabilities other than these, a method due to Anderson (1972) may be employed. If $\hat{q}_1^*$ and $\hat{q}_2^*$ are the desired estimates of the prior probabilities, then the correction term $\ln(n_2\hat{q}_1^* / n_1\hat{q}_2^*)$ is added to the logistic regression's estimate of α_0. Using this correction term, the linearized classification rules estimated by logistic regression and linear discrimination are completely analagous, and the NS method of error rate estimation may be applied similarly.

In quadratic discrimination, the covariance matrices for the two populations are estimated separately (denote these estimates by $\underset{\sim}{S}_1$ and $\underset{\sim}{S}_2$) and are not assumed to be equal. The QDF is of the form

$$Q(\underset{\sim}{x}) = a_0^* + \underset{\sim}{x}'\underset{\sim}{a}^* + \underset{\sim}{x}'\underset{\sim}{A}^*\underset{\sim}{x},$$

where $a_0^* = -\hat{c} + \frac{1}{2}[\ \ln|\underset{\sim}{S}_2| - \ln|\underset{\sim}{S}_1|\ ] - \frac{1}{2}(\bar{\underset{\sim}{x}}_1' \underset{\sim}{S}_1^{-1} \bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2' \underset{\sim}{S}_2^{-1} \bar{\underset{\sim}{x}}_2)$, $\underset{\sim}{a}^* = \underset{\sim}{S}_1^{-1} \bar{\underset{\sim}{x}}_1 - \underset{\sim}{S}_2^{-1} \bar{\underset{\sim}{x}}_2$, and $\underset{\sim}{A}^* = \frac{1}{2}(\underset{\sim}{S}_2^{-1} - \underset{\sim}{S}_1^{-1})$. In order to apply the NS method in this case, the LDF may simply be replaced by $Q(\underset{\sim}{x})$.

The location model is sometimes used when the explanatory variables consist of a mixture of r_1 continuous variables and r_2 categorical variables. Although a modification by Krzanowski (1975) shows how the location model can be altered to reduce the number of parameters estimated, this method essentially consists of a separate LDF using the r_1 continuous variables as the independent variables within each of the cells defined by the r_2 categorical variables. For example, in the data set considered here, there are three binary variables, and therefore eight distinct linear discriminant functions are estimated, each with z_1 and z_2 as independent variables. Within each cell, indexed by i, let n_{i1} and n_{i2} be the training sample sizes; let $\underset{\sim}{x}_{ij}$, $j=1,\ldots,n_{i1}$, be the observations in the training sample from π_1; let m_i^2 be the estimated Mahalanobis distance between the populations; let $b_i = \{ [(r_1+2)(n_{i1}-1) + (n_{i2}-1)] / [n_{i1}(n_{i1}+n_{i2}-r_1-3)] \}^{\frac{1}{2}}$ be the smoothing constant; and let $w_i(\underset{\sim}{x})$ be the linear discriminant function. The NS method applied to the location model can be defined as

$$\hat{\alpha} = n_1^{-1} \sum_i \sum_{j=1}^{n_{i1}} \Phi\{ [\hat{c}-w_i(\underset{\sim}{x}_{ij})] / b_i m_i \}.$$

A FORTRAN SUBROUTINE which computes the NS estimate is given in Appendix 2. In order to use this subprogram, the appropriate discrimination procedure must already have been performed. The training samples, training sample sizes, number of dimensions in the observation vector, and the square root of the Mahalanobis distance between the populations must all be entered as parameters into this subprogram. In addition, the discriminant function, regardless of the type of discriminant procedure used, must be supplied in a separate FUNCTION subprogram. Two examples of this subprogram are given in Appendix 2. Two NS error rate estimates are returned by the SUBROUTINE: one for observations from population 1 and another for observations from population 2.

When the location model is used, the FORTRAN SUBROUTINE must be run separately for each of the cells defined by the r_2 categorical variables. For each cell, indexed by i, a separate set of parameters must be entered into the SUBROUTINE, and a separate FUNCTION subprogram with the appropriate discriminant function must be written. Error rates for cell i and population j, j=1,2, labelled $\hat{\alpha}_{ij}$, are obtained. In order to obtain overall estimates of the error rates,

$\hat{\alpha}_1$ and $\hat{\alpha}_2$, a weighted average must be taken:

$$\hat{\alpha}_j = n_j^{-1} \sum_i n_{ij} \hat{\alpha}_{ij}, \quad j=1,2.$$

This is not done by the FORTRAN SUBROUTINE.

The discriminant rules considered by Knoke (1982) include the following: LDF (linear discrimination using all five independent variables); LR: LDF (logistic regression using all five independent variables), LDF + HT (linear discrimination using z_1, z_2, and $y_1y_2y_3$); LR: LDF + HT (logistic regression using z_1, z_2, and $y_1y_2y_3$); LR: LDF + HHT (logistic regression using z_1, z_2, y_1, y_2, y_3, y_1y_2, y_1y_3, and y_2y_3); QDF (quadratic discrimination using all five independent variables); and LM (location model using all five independent variables). Each of these seven rules is considered twice: once with $\hat{c} = -1.405$, which corresponds to the prior probabilities estimated by the training sample sizes (n_1=110, n_2=27), and called the minimum misclassifications rule; and again with $\hat{c} = 0$, which corresponds to equal prior probabilities, and called the minimax rule.

The results of these analyses are presented in Table 5.1. The columns labelled R correspond to the resubstitution estimates, NS to the NS estimates, U to the leave-one-out estimates, and I to the error rate estimates based on the independent testing samples. For the minimax rule, the average error rate is the simple average of the two values; while for the minimum misclassifications rule, the average error rate is a weighted average of the two values, where the weights correspond to the appropriate sample sizes.

The main purpose of this chapter is to illustrate how the NS method can be used in a practical situation, and this has been done.

TABLE 5.1

ERROR RATE ESTIMATES FOR SEVERAL DISCRIMINANT RULES:

FOR PATIENTS WHO LIVED (L), THOSE WHO DIED (D),

AND THE AVERAGE (A)

		Minimum Misclassifications Rule				Minimax Rule			
		R	NS	U	I	R	NS	U	I
LDF	L	5	5	6	8	26	25	27	22
	D	56	58	59	57	22	24	22	29
	A	15	15	17	14	24	24	25	25
LR: LDF	L	5	5	-	7	25	25	-	22
	D	56	57	-	71	22	24	-	29
	A	15	16	-	15	24	24	-	25
LDF + HT	L	4	4	4	5	14	17	14	14
	D	52	52	52	64	26	31	30	29
	A	13	14	13	13	20	24	22	21
LR: LDF + HT	L	4	5	-	5	18	21	-	15
	D	56	55	-	64	19	25	-	29
	A	14	15	-	13	18	23	-	22
LR: LDF + HHT	L	5	5	-	7	20	22	-	16
	D	56	55	-	64	19	25	-	29
	A	15	15	-	14	19	23	-	23
QDF	L	5	5	5	8	12	13	15	12
	D	56	58	63	43	37	37	44	21
	A	15	15	17	12	24	25	29	17
LM	L	5	6	6	8	16	18	18	16
	D	48	47	63	50	22	28	48	14
	A	13	14	18	13	19	23	33	15

A few other comments can also be made. The NS estimates in Table 5.1 seem to be reasonable in that they are fairly close to the other estimates; and in general the NS estimates tend to be intermediate in value between the resubstitution estimate, which is usually the smallest, and the leave-one-out estimate, which is usually the largest. This, of course, is the rank order that would be expected based on the relative biases of the estimators: the R method has a large optimistic bias, the NS method tends to have a smaller optimistic bias, and the U method is nearly unbiased. Notice that these three estimates tend to be relatively similar for the group of patients who lived, but are sometimes very dissimilar for the group of patients who died. This reflects that the training sample for the former group is larger than for the latter, and that all three estimators asymptotically converge on the actual error rate.

VI. CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER RESEARCH

Under the criterion of minimizing the UMSE, the NS method is never much worse than any other error rate estimator, but with some types of non-normal parent distributions it can be considerably better than all other estimators. In addition, smoothing seems to be of most value in exactly the situations where it is most necessary. Therefore, smoothing should certainly be considered as an alternative to all other estimators.

In some situations, error rate estimators other than the NS method are recommended for use. If more than two independent variables are used in the discriminant analysis, and if the Mahalanobis distance between the parent distributions is felt to be relatively small, then the leave-one-out estimator should be chosen. If this is not the case, and if one is reasonably certain that the parent distributions are normal, then the DS method is probably the best choice. However, if a single estimator is to be chosen for use in all situations, then, based on the results of this study, this estimator should be the NS method.

While the NS method appears to be a good smoothed estimator, it is, however, not necessarily the best one. In a future study, two aspects of smoothing might be examined in order to improve upon the NS method. Very little difference was found between the two types of smoothing functions examined in this paper. However, it is certainly possible that some other smoothing function will prove to have advantages over both normal smoothing and the smooth slide. Assuming that normal smoothing is used, the value of b given by (4.6) does not necessarily define in any sense the optimum level of smoothing. While this equation has the advantage of being based on theoretical, rather than ad hoc, considerations, another equation might be found which yields an estimator with consistently lower UMSE than that of the NS method.

When the amount of smoothing is substantial, the UMSEs of the smoothed estimators were shown to be very sensitive to the type of parent distribution. While the NS method does not appear to suffer from a lack of robustness in any of the situations studied, this potential problem should be examined more closely. In particular, the level of smoothing defined by (4.6) increases with the number of independent variables: Does this have any effect on the robustness of the NS method? To help answer this question, several type of non-normal parent distributions should be studied, both in the univariate and multivariate cases. For the same reasons, the NS method should also be examined in other real data situations, such as the one presented in Chapter V.

In that example, the NS method was used in situations where discriminant functions other than the LDF were being evaluated. While the application of the NS method in these situations seemed reasonable, the theoretical development of this method was restricted to the LDF. It is therefore important to examine the quality of the NS method as measured by UMSE for logistic regression, quadratic discrimination, and the location model.

The bootstrap is a statistical technique which can be applied to the estimation of error rates in discriminant analysis. As discussed in Section I.2.8, this technique can be used to reduce the bias of the resubstitution estimator relative to the expected actual error rate; however, there is nothing to prevent one from similarly using the bootstrap to reduce the bias of the NS method. Since the major advantage of smoothing is a variance reduction, and that of the bootstrap is a bias reduction, application of the bootstrap to the NS method might be a valuable complement. This possibility should be explored.

In conclusion, the results obtained in this paper on smoothing and the NS method certainly seem to be encouraging enough to warrant further research.

A.1 AN EXPRESSION FOR E[h(U)h(V)] IN THE UNIVARIATE NORMAL CASE FOR THE SMOOTH SLIDE ESTIMATOR

$$
\begin{aligned}
E[h(U)h(V)] = & \int_{-\infty}^{d-\beta} \int_{-\infty}^{d-\beta} g(y,z;\rho)\, dy\, dz \\
& + \int_{-\infty}^{d-\beta} \int_{d-\beta}^{d+\beta} [(d+\beta-y)/2\beta]\, g(y,z;\rho)\, dy\, dz \\
& + \int_{d-\beta}^{d+\beta} \int_{-\infty}^{d-\beta} [(d+\beta-z)/2\beta]\, g(y,z;\rho)\, dy\, dz \\
& + \int_{d-\beta}^{d+\beta} \int_{d-\beta}^{d+\beta} [(d+\beta-y)/2\beta]\, [(d+\beta-z)/2\beta]\, g(y,z;\rho)\, dy\, dz \\
= & \int_{-\infty}^{d-\beta} \int_{-\infty}^{d-\beta} g(y,z;\rho)\, dy\, dz \\
& + [(d+\beta)/2\beta] \int_{-\infty}^{d-\beta} \int_{d-\beta}^{d+\beta} g(y,z;\rho)\, dy\, dz \\
& + [(d+\beta)/2\beta] \int_{d-\beta}^{d+\beta} \int_{-\infty}^{d-\beta} g(y,z;\rho)\, dy\, dz \\
& + [(d+\beta)/2\beta]^2 \int_{d-\beta}^{d+\beta} \int_{d-\beta}^{d+\beta} g(y,z;\rho)\, dy\, dz \\
& + (2\beta)^{-1} \int_{-\infty}^{d-\beta} \int_{d-\beta}^{d+\beta} (-y)\, g(y,z;\rho)\, dy\, dz \\
& + (2\beta)^{-1} \int_{d-\beta}^{d+\beta} \int_{-\infty}^{d-\beta} (-z)\, g(y,z;\rho)\, dy\, dz \\
& + [(d+\beta)/4\beta^2] \int_{d-\beta}^{d+\beta} \int_{d-\beta}^{d+\beta} (-y)\, g(y,z;\rho)\, dy\, dz \\
& + [(d+\beta)/4\beta^2] \int_{d-\beta}^{d+\beta} \int_{d-\beta}^{d+\beta} (-z)\, g(y,z;\rho)\, dy\, dz \\
& + (2\beta)^{-2} \int_{d-\beta}^{d+\beta} \int_{d-\beta}^{d+\beta} (yz)\, g(y,z;\rho)\, dy\, dz \qquad \text{(A.1)}
\end{aligned}
$$

This result consists of the sum of nine double integrals. The first four integrals are evaluated using the result

$$\int_{-\infty}^{k_1} \int_{-\infty}^{k_2} g(y,z;\rho)\, dy\, dz = G(k_1,k_2;\rho). \qquad \text{(A.2)}$$

In order to evaluate the other five integrals, the following results are first derived.

Let $S = Y - \rho Z$ and $T = Z$.

$$\int_{-\infty}^{k_1}\int_{-\infty}^{k_2} (-y+\rho z)\ g(y,z;\rho)\ dy\ dz$$

$$= \int_{-\infty}^{k_1}\int_{-\infty}^{k_2} (-y+\rho z)\ [2\pi(1-\rho^2)^{\frac{1}{2}}]^{-1} \exp[-(y-\rho z)^2/2(1-\rho^2)]\ \exp(-z^2/2)\ dy\ dz$$

$$= \int_{-\infty}^{k_1}\int_{-\infty}^{k_2-\rho t} [2\pi(1-\rho^2)^{\frac{1}{2}}]^{-1} \exp[-s^2/2(1-\rho^2)]\ \exp(-t^2/2)\ (-s ds)\ dt$$

$$= \int_{-\infty}^{k_1} [(1-\rho^2)^{\frac{1}{2}}/2\pi]\ \exp[-(k_2-\rho t)^2/2(1-\rho^2)]\ \exp(-t^2/2)\ dt$$

$$= \int_{-\infty}^{k_1} [(1-\rho^2)^{\frac{1}{2}}/2\pi]\ \exp(-k_2{}^2/2)\ \exp[-(t-k_2\rho)^2/2(1-\rho^2)]\ dt$$

$$= (1-\rho^2)\ \phi(k_2)\ \Phi[(k_1-k_2\rho)(1-\rho^2)^{-\frac{1}{2}}]. \qquad (A.3)$$

Similarly,

$$\int_{-\infty}^{k_1}\int_{-\infty}^{k_2} (-z+\rho y)\ g(y,z;\rho)\ dy\ dz = (1-\rho^2)\ \phi(k_1)\ \Phi[(k_2-k_1\rho)(1-\rho^2)^{-\frac{1}{2}}]. \qquad (A.4)$$

$$\int_{-\infty}^{k_1}\int_{-\infty}^{k_2} (y-\rho z)^2\ g(y,z;\rho)\ dy\ dz$$

$$= \int_{-\infty}^{k_1}\int_{-\infty}^{k_2-\rho t} [2\pi(1-\rho^2)^{\frac{1}{2}}]^{-1} \exp[-s^2/2(1-\rho^2)]\ \exp(-t^2/2)\ s^2\ ds\ dt$$

$$= \int_{-\infty}^{k_1}\int_{-\infty}^{k_2-\rho t} [(1-\rho^2)^{\frac{1}{2}}/2\pi]\ \exp[-s^2/2(1-\rho^2)]\ \exp(-t^2/2)\ ds\ dt$$

$$- \int_{-\infty}^{k_1} [(1-\rho^2)^{\frac{1}{2}}/2\pi](k_2-\rho t)\ \exp[-(k_2-\rho t)^2/2(1-\rho^2)]\ \exp(-t^2/2)\ dt$$

$$= \int_{-\infty}^{k_1}\int_{-\infty}^{k_2} [(1-\rho^2)^{\frac{1}{2}}/2\pi] \exp[-(y-\rho z)^2/2(1-\rho^2)] \exp(-z^2/2)\ dy\ dz$$

$$- \rho\int_{-\infty}^{k_1} [(1-\rho^2)^{\frac{1}{2}}/2\pi]\exp(-k_2^2/2)\ \exp[-(t-\rho k_2)^2/2(1-\rho^2)]\ (-t+\rho k_2)\ dt$$

$$- k_2(1-\rho^2)\int_{-\infty}^{k_1} [(1-\rho^2)^{\frac{1}{2}}/2\pi]\exp(-k_2^2/2)\ \exp[-(t-\rho k_2)^2/2(1-\rho^2)]\ dt$$

$$= (1-\rho^2)\ G(k_1,k_2;\rho) - \rho(1-\rho^2)^2\ g(k_1,k_2;\rho)$$

$$- k_2(1-\rho^2)^2\ \phi(k_2)\ \Phi[(k_1-\rho k_2)(1-\rho^2)^{-\frac{1}{2}}]. \qquad (A.5)$$

Similarly,

$$\int_{-\infty}^{k_1}\int_{-\infty}^{k_2} (z-\rho y)^2\ g(y,z;\rho)\ dy\ dz = (1-\rho^2)\ G(k_1,k_2;\rho)$$

$$- \rho(1-\rho^2)^2\ g(k_1,k_2;\rho) - k_1(1-\rho^2)^2\ \phi(k_1)\ \Phi[(k_2-k_1\rho)(1-\rho^2)^{-\frac{1}{2}}]. \qquad (A.6)$$

$$\int_{-\infty}^{k_1}\int_{-\infty}^{k_2} (y-\rho z)\ (z-\rho k_2)\ g(y,z;\rho)\ dy\ dz$$

$$= \int_{-\infty}^{k_1}\int_{-\infty}^{k_2-\rho t} [2\pi(1-\rho^2)^{\frac{1}{2}}]^{-1} \exp[-s^2/2(1-\rho^2)]\ \exp(-t^2/2)\ s\ (t-\rho k_2)\ ds\ dt$$

$$= -\int_{-\infty}^{k_1} [(1-\rho^2)^{\frac{1}{2}}/2\pi]\ \exp[-(k_2-\rho t)^2/2(1-\rho^2)]\ \exp(-t^2/2)\ (t-\rho k_2)\ dt$$

$$= -\int_{-\infty}^{k_1} [(1-\rho^2)^{\frac{1}{2}}/2\pi]\ \exp(-k_2^2/2)\ \exp[-(t-\rho k_2)^2/2(1-\rho^2)]\ (t-\rho k_2)\ dt$$

$$= (1-\rho^2)^{3/2}(2\pi)^{-1}\ \exp(-k_2^2/2)\ \exp[-(k_1-\rho k_2)^2/2(1-\rho^2)]$$

$$= (1-\rho^2)^2\ g(k_1,k_2;\rho). \qquad (A.7)$$

Similarly,

$$\int_{-\infty}^{k_1}\int_{-\infty}^{k_2} (z-\rho y)(y-\rho k_1)\ g(y,z;\rho)\ dy\ dz = (1-\rho^2)^2\ g(k_1,k_2;\rho). \quad (A.8)$$

Since $-y = [(-y+\rho z) + \rho(-z+\rho y)] \ / \ (1-\rho^2)$, combine (A.3) and (A.4) to obtain

$$\int_{-\infty}^{k_1}\int_{-\infty}^{k_2} (-y)\ g(y,z;\rho)\ dy\ dz = \phi(k_2)\ \Phi[(k_1-\rho k_2)(1-\rho^2)^{-\frac{1}{2}}]$$

$$+ \rho\ \phi(k_1)\ \Phi[(k_2-\rho k_1)(1-\rho^2)^{-\frac{1}{2}}]. \qquad (A.9)$$

And finally, since

$$yz = \rho(1+\rho^2)\ \{\ [(z-\rho y)(y-\rho k_1) + (y-\rho z)(z-\rho k_2)]\ /\ \rho$$

$$+ [(y-\rho z)^2 + (z-\rho y)^2]\ /\ (1+\rho^2)$$

$$- k_2(-y+\rho z) - k_1(-z+\rho y)\ \}\ /\ 2(1-\rho^2),$$

(A.3) through (A.8) are combined, and after much algebra the following result is obtained.

$$\int_{-\infty}^{k_1}\int_{-\infty}^{k_2} (yz)\ g(y,z;\rho)\ dy\ dz = \rho G(k_1,k_2;\rho) + (1-\rho^2)\ g(k_1,k_2;\rho)$$

$$- \rho k_1\ \phi(k_1)\ \Phi[(k_2-\rho k_1)(1-\rho^2)^{-\frac{1}{2}}] - \rho k_2\ \phi(k_2)\ \Phi[(k_1-\rho k_2)(1-\rho^2)^{-\frac{1}{2}}].$$

$$(A.10)$$

The results (A.2), (A.9) and (A.10) are substituted into (A.1), and again, after much algebra:

$$
\begin{aligned}
E[h(U)h(V)] = \{ & [(d+\beta)^2+\rho]\ G(d+\beta,d+\beta;\rho) - 2[(d+\beta)(d-\beta)+\rho]\ G(d-\beta,d+\beta;\rho) \\
& + [(d-\beta)^2+\rho]\ G(d-\beta,d-\beta;\rho) + (1-\rho^2)\ g(d+\beta,d+\beta;\rho) \\
& - 2(1-\rho^2)\ g(d-\beta,d+\beta;\rho) + (1-\rho^2)\ g(d-\beta,d-\beta;\rho) \\
& + 2(d+\beta)\ \phi(d+\beta)\ \Phi[\ (d+\beta-\rho d-\rho\beta)\ (1-\rho^2)^{-\frac{1}{2}}\ ] \\
& - 2(d-\beta)\ \phi(d+\beta)\ \Phi[\ (d-\beta-\rho d-\rho\beta)\ (1-\rho^2)^{-\frac{1}{2}}\ ] \\
& - 2(d+\beta)\ \phi(d-\beta)\ \Phi[\ (d+\beta-\rho d+\rho\beta)\ (1-\rho^2)^{-\frac{1}{2}}\ ] \\
& + 2(d-\beta)\ \phi(d-\beta)\ \Phi[\ (d-\beta-\rho d+\rho\beta)\ (1-\rho^2)^{-\frac{1}{2}}\ ]\ \}\ /\ 4\beta^2.
\end{aligned}
$$

A.2 A FORTRAN SUBROUTINE TO EVALUATE THE NS ESTIMATE

```
      SUBROUTINE SMOOTH(DATA,N,NDIM,M,NS)
      DIMENSION DATA(2,10,200),N(2),B(2),X(10)
      REAL M,NS(2)
      B(1) = SQRT(FLOAT((NDIM+2)*(N(1)-1) + (N(2)-1)) / FLOAT(N(1)*
     X (N(1)+N(2)-NDIM-3)))
C
C  THIS IS THE SMOOTHING CONSTANT FOR THE ESTIMATOR OF ERROR RATES
C  FOR POPULATION 1.
C
      B(2) = SQRT(FLOAT((NDIM+2)*(N(2)-1) + (N(1)-1)) / FLOAT(N(2)*
     X (N(1)+N(2)-NDIM-3)))
C
C  THIS IS THE SMOOTHING CONSTANT FOR THE ESTIMATOR OF ERROR RATES
C  FOR POPULATION 2.
C
      DO 10 I=1,2
      NI=N(I)
      NS(I)=0.
C
C  A SEPARATE ESTIMATE (NS(I)) IS OBTAINED FOR EACH POPULATION,
C  I=1 AND I=2.
C
      DO 10 J=1,NI
C
C  SUM OVER EACH SAMPLE OBSERVATION FROM POPULATION I.
C
      DO 20 K=1,NDIM
   20 X(K)=DATA(I,K,J)
C
C  SET THE VECTOR X EQUAL TO THE NEXT SAMPLE OBSERVATION.
C
      Z=F(X)/(B(I)*M)
      IF (I .EQ. 1) Z = -1.*Z
      CALL MDNOR(Z,A)
C
C  THE VARIABLE A TAKES ON THE VALUE OF THE SMOOTHING FUNCTION.
C  THE FUNCTION SUBROUTINE F(X) MUST BE WRITTEN BY THE USER, AND THE
C  SUBROUTINE MDNOR IS AVAILABLE IN IMSL.
C
   10 NS(I) = NS(I) + A/FLOAT(N(I))
C
C  THE ESTIMATE IS THE MEAN VALUE OF A.
C
      RETURN
      END
```

Variables Entered Into SUBROUTINE SMOOTH

DATA - The matrix DATA(I,J,K) contains the training samples from group 1 (I=1) and group 2 (I=2). Space is available for as many as ten independent variables (J) and two hundred observations per training sample (K).

N - N(1) is the number of observations from group 1 and N(2) is the number of observations from group 2.

NDIM - The number of independent variables.

M - The square root of the distance between the sample means, weighted by the inverse of the pooled covariance matrix.

NS - NS(1) and NS(2) are the smoothed error rate estimates for populations 1 and 2 which are returned to the main program.

Subprograms Called by SUBROUTINE SMOOTH

MDNOR(Z,A) - This subroutine, available in the IMSL library, returns A such that A = Φ(Z). Similar subroutines are readily available in other statistical packages.

F(X) - This function subroutine, which must be supplied by the user, evaluates the discriminant function. Two examples of this subprogram follow.

```
 FUNCTION F(X)
 DIMENSION X(10)
 W = -1.437 + 0.02323*X(1) + 0.05636*X(2) + 0.79179*X(3)
 RETURN
 END

 FUNCTION F(X)
 DIMENSION X(10)
 W = 1.247 + 0.041*X(1) - 0.184*X(2) - 1.118*X(1)*X(1)
X  + 0.823*X(2)*X(2) + 1.124*X(1)*X(2)
 RETURN
 END
```

BIBLIOGRAPHY

ANDERSON, J.A. (1972), "Separate Sample Logistic Discrimination," Biometrika, 59, 19-35.

ANDERSON, T.W. (1951), "Classification by Multivariate Analysis," Psychometrika, 16, 31-50.

BARTLETT, M.S. (1951), "An Inverse Matrix Adjustment Arising in Discriminant Analysis," Annals of Mathematical Statistics, 22, 107-111.

BRAILOVSKIY, V.L. (1964), "An Object Recognition Algorithm with Many Parameters and its Applications," Engineering Cybernetics, 2, 22-30.

BRAILOVSKIY, V.L., and LUNTS, A.L. (1964), "The Multiparameter Recognition Problem and its Solution," Engineering Cybernetics, 1, 13-22.

BROFFITT, J.D., and WILLIAMS, J.S. (1973), "Minimum Variance Estimators for Misclassification Probabilities in Discriminant Analysis," Journal of Multivariate Analysis, 3, 311-327.

CARNAHAN, B., LUTHER, H.A., and WILKES, J.O. (1969), Numerical Methods, New York: John Wiley.

CHEN, E.H. (1971), "A Random Normal Number Generator for 32-Bit-Word Computers," Journal of the American Statistical Association, 66, 400-403.

DAS GUPTA, S. (1974), "Probability Inequalities and Errors in Classification," Annals of Statistics, 2, 751-762.

DUDA, R.O., and HART, P.E. (1973), Pattern Classification and Scene Analysis, New York: John Wiley.

DUNN, O.J. (1971), "Some Expected Values for Probabilities of Correct Classification in Discriminant Analysis," Technometrics, 13, 345-353.

DUNN, O.J., and VARADY, P.D. (1966), "Probabilities of Correct Classification in Discriminant Analysis," Biometrics, 22, 908-924.

EFRON, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," Annals of Statistics, 7, 1-26.

EFRON, B., and GONG, G. (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," The American Statistician, 37, 36-48.

FISHER, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, 7, 179-188.

FUKUNAGA, K., and KESSELL, D.L. (1971), "Estimation of Classification Error," IEEE Transactions on Computers, C-20, 1521-1527.

GANESALINGAM, S., and McLACHLAN, G.J. (1980), "Error Rate Estimation on the Basis of Posterior Probabilities," Pattern Recognition, 12, 405-413.

GLICK, N. (1978), "Additive Estimators for Probabilities of Correct Classification," Pattern Recognition, 10, 211-222.

HAMMERSLEY, J.M., and HANDSCOMB, D.C. (1964), Monte Carlo Methods, London: Methuen & Co., Ltd.

HILLS, M. (1966), "Allocation Rules and Their Error Rates," Journal of the Royal Statistical Society, Ser. B, 28, 1-20.

INTERNATIONAL MATHEMATICAL AND STATISTICAL LIBRARIES (1979), IMSL Library Reference Manual (Ed. 7), Houston: IMSL.

JENNRICH, R., and SAMPSON, P. (1981), "Stepwise Discriminant Analysis," in BMDP Statistical Software, ed. W.J. Dixon, Los Angeles: University of California Press, 519-537.

JOHN, S. (1961), "Errors in Discrimination," Annals of Mathematical Statistics, 32, 1125-1144.

KNOKE, J.D. (1982), "Discriminant Analysis with Discrete and Continuous Variables," Biometrics, 38, 191-200.

KRZANOWSKI, W.J. (1975), "Discrimination and Classification Using Both Binary and Continuous Variables," Journal of the American Statistical Association, 70, 782-790.

LACHENBRUCH, P.A. (1967), "An Almost Unbiased Method of Obtaining Confidence Intervals for the Probability of Misclassification in Discriminant Analysis," Biometrics, 23, 639-645.

__________ (1968), "On Expected Probabilities of Misclassification in Discriminant Analysis, Necessary Sample Size, and a Relation with the Multiple Correlation Coefficient," Biometrics, 24, 823-834.

LACHENBRUCH, P.A., and MICKEY, M.R. (1968), "Estimation of Error Rates in Discriminant Analysis," Technometrics, 10, 1-11.

LURIA, M.H., KNOKE, J.D., MARGOLIS, R.M., HENDRICKS, F.H., and KUPLIC, J.B. (1976), "Acute Myocardial Infarction: Prognosis After Recovery," Annals of Internal Medicine, 85, 561-565.

McLACHLAN, G.J. (1972), "An Asymptotic Expansion for the Variance of the Errors of Misclassification of the Linear Discriminant Function," Australian Journal of Statistics, 14, 68-72.

__________ (1973), "An Asymptotic Expansion of the Expectation of the Estimated Error Rate in Discriminant Analysis, Australian Journal of Statistics, 15, 210-214.

__________ (1974a), "The Asymptotic Distributions of the Conditional Error Rate and Risk in Discriminant Analysis," Biometrika, 61, 131-137.

__________ (1974b), "Estimation of the Errors of Misclassification on the Criterion of Asymptotic Mean Squared Error," Technometrics, 16, 255-260.

__________ (1974c), "An Asymptotic Unbiased Technique for Estimating the Error Rates in Discriminant Analysis," Biometrics, 30, 239-249.

__________ (1974d), "The Relationship in Terms of Asymptotic Mean Squared Error Between the Separate Problems of Estimating Each of Three Types of Error Rate of the Linear Discriminant Function," Technometrics, 16, 569-575.

__________ (1975), "Confidence Intervals for the Conditional Probability of Misallocation in Discriminant Analysis," Biometrics, 31, 161-167.

__________ (1976), "The Bias of the Apparent Error Rate in Discriminant Analysis," Biometrika, 63, 239-244.

__________ (1980), "The Efficiency of Efron's 'Bootstrap' Approach Applied to Error Rate Estimation in Discriminant Analysis," Journal of Statistical Computation and Simulation, 11, 273-279.

MOORE, D.S., WHITSITT, S.J., and LANDGREBE, D.A. (1976), "Variance Comparisons for Unbiased Estimators of Probability of Correct Classification," IEEE Transactions on Information Theory, 22, 102-105.

MORAN, M.A. (1975), "On the Estimation of Errors of Allocation Associated with a Linear Discriminant Function," Biometrika, 62, 141-148.

NEAVE, H.R. (1975), "On Using the Box-Muller Transformation with Multiplicative Congruential Pseudo-random Number Generators," Applied Statistics, JRSS Series C, 22, 92-97.

OKAMOTO, M. (1963), "An Asymptotic Expansion for the Distribution of the Linear Discriminant Function," Annals of Mathematical Statistics, 34, 1286-1301.

QUENOUILLE, M.H. (1956), "Notes on Bias in Estimation," Biometrika, 43, 353-360.

SAYRE, J.W. (1980), "The Distributions of the Actual Error Rates in Linear Discriminant Analysis," Journal of the American Statistical Association, 75, 201-205.

SCHERVISH, M.J. (1981), "Asymptotic Expansions for Correct Classification Rates in Discriminant Analysis," Annals of Statistics, 9, 1002-1009.

SEDRANSK, N., and OKAMOTO, M. (1971), "Estimation of the Probabilities of Misclassification for a Linear Discriminant Function in the Univariate Normal Case," Annals of the Institute of Statistical Mathematics, 23, 419-435.

SMITH, C.A.B. (1947), "Some Examples of Discrimination," Annals of Eugenics, 18, 272-282.

SNEDECOR, G.W., and COCHRAN, W.C. (1967), Statistical Methods (Ed. 6), Ames, Iowa: The Iowa State University Press.

SORUM, M.J. (1971), "Estimating the Conditional Probability of Misclassification," Technometrics, 13, 333-343.

__________ (1972), "Estimating the Expected and the Optimal Probabilities of Misclassification," Technometrics, 14, 935-943.

__________ (1973), "Estimating the Expected Probability of Misclassification for a Rule Based on the Linear Discriminant Function: Univariate Normal Case," Technometrics, 15, 329-339.

TOUSSAINT, G.T. (1974), "Bibliography on Estimation of Misclassification," IEEE Transactions on Information Theory, IT-20, 472-479.

WALD, A. (1944), "On a Statistical Problem Arising in the Classification of an Individual into One of Two Groups," Annals of Mathematical Statistics, 15, 145-162.