VARIANCE COMPARISONS FOR UNBIASED ESTIMATORS OF PROBABILITIES OF CORRECT CLASSIFICATIONS

by

David S. Moore
Stephen J. Whitsitt
David A. Landgrebe

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #401

January 1975

Research sponsored in part by the Air Force Office of Scientific Research, Air Force Systems Command, under Grant No. AFOSR-72-2350B and in part by NASA Grant NGL 15-005-112.

D. S. Moore is with the Department of Statistics, Purdue University West Lafayette, Indiana, 47907.

S. J. Whitsitt was with the Laboratory for Applications of Remote Sensing. He is now with the Data Systems Department, TRW Systems Group, One Space Park, Redondo Reach, California, 90278.

D. A. Landgrebe is with the Department of Electrical Engineering and the Laboratory for Applications of Remote Sensing (LARS), Purdue University, West Lafayette, Indiana, 47907.

VARIANCE COMPARISONS FOR UNBIASED ESTIMATORS OF PROBABILITIES OF CORRECT CLASSIFICATIONS

David S. Moore
Stephen J. Whitsitt
David A. Landgrebe

Abstract

Variance relationships among certain count estimators and posterior-probability estimators of correct recognition are investigated. A statistic using posterior probabilities is presented for use in stratified sampling designs. A test case involving three normal classes is examined.

Research sponsored in part by the Air Force Office of Scientific Research, Air Force Systems Command, under Grant No. AFOSR-72-2350B and in part by NASA Grant NGL 15-005-112.

D. S. Moore is with the Department of Statistics, Purdue University, West Lafayette, Indiana, 47907.

S. J. Whitsitt was with the Laboratory for Applications of Remote Sensing. He is now with the Data Systems Department, TRW Systems Group, One Space Park, Redondo Beach, California, 90278.

D. A. Landgrebe is with the Department of Electrical Engineering and the Laboratory for Applications of Remote Sensing (LARS), Purdue University, West Lafayette, Indiana, 47907.

I. INTRODUCTION

Let X be an observation (possibly multivariate) which is to be classified into one of M classes $\omega_1,\ldots,\omega_M$. Suppose further that $P_1,\ldots,P_M$ are the prior probabilities of classes $\omega_1,\ldots,\omega_M$ and that $p_i(x)$ is the probability density function of X given that it belongs to class ω_i. Unclassified observations then have the mixture density

$$p(x) = \sum_{i=1}^{M} P_i p_i(x). \tag{1}$$

An arbitrary classification rule may be described as follows: classify X as belonging to class ω_i if X falls in Γ_i, where $\Gamma_1,\ldots,\Gamma_M$ are sets which partition the observation space. (We do not here consider rules which allow refusing to classify.) Let $I_i(x)$ be the indicator function of Γ_i,

$$\begin{aligned} I_i(x) &= 1 \qquad x \in \Gamma_i, \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

Then the probability of correct classification for an X from class ω_i is

$$p_{ci} = \int_{\Gamma_i} p_i(x)dx = \int I_i(x)p_i(x)dx \tag{2}$$

and the probability of correct classification for an unclassified observation X is

$$p_c = \sum_{i=i}^{M} P_i p_{ci}. \tag{3}$$

Estimation of p_c (or equivalently, of the probability of error, $1 - p_c$) from sample data is of considerable importance in situations where direct calculation of the p_{ci} is difficult and Monte Carlo methods must be used. Two familiar methods for estimating p_c are random sampling and selective (stratified) sampling [1, Sec. 5.4; 2; 3, p. 255]. In both methods, the statistic for error is based on the number of correctly classified samples. In the case of selective sampling, however, the number of samples used to estimate error from each class must be within the control of the statistician, and the prior probabilities P_i must be known. In this sense, it is sometimes said that the first method employs unclassified samples, and the second method, classified samples. This distinction is somewhat academic in view of the fact that the true classification of each sample must ultimately be known in order to determine the number of misclassifications.

In a different approach to estimating p_c, Fukunaga and Kessell [4] extended the use of the reject function of Chow [5] to an unbiased statistic for error by using the posterior probabilities $p(\omega_i|X)$ at each sample X. However, knowledge of the priors P_i and class densities $p_i(x)$ is required, although estimates of these quantities for a specific recognition problem have been employed with apparently good results by the same authors [6]. This method uses unclassified samples from the mixture density, as does random sampling (although, unlike random sampling, the class assignments are never required).

The relationship between the posterior random sampling statistic of Fukunaga and Kessell and selective sampling deserves some attention. In many situations where Monte Carlo procedures are usually required for estimating p_c (arbitrary Gaussian ω_i, for instance), the statistician can control the selection of the samples. An example might be computer simulation. In such cases, the requirement that each sample come from the mixture density may require more computation, since this implies that one must randomize, according to the priors, on the class labels ω_i. Also, selective sampling results in variance no larger than random sampling [2], as does the posterior random sampling statistic [4]. But the relationship between the former and the latter has not been established.

Since it is economically desirable to use unbiased statistics for

p_c with minimum variance, we will examine the variance relationships among several statistics employing both classified and unclassified samples. A new statistic for p_c which uses both posterior probabilities and class assignments will be introduced.

II. COUNT ESTIMATORS

The standard estimator of a probability is simply the proportion of observations falling in the event in question. Suppose then that $X_1,\ldots,X_N$ are unclassified samples, i.e., independent random vectors each distributed according to $p(x)$. The proportion of correct classifications $\hat{p}_1$ is an unbiased estimator of p_c having variance

$$\sigma^2(\hat{p}_1) = \frac{1}{N}(p_c - p_c^2). \tag{4}$$

In planning a simulation experiment, we can choose instead to distribute N observations among the classes, taking N_i observations from class ω_i, where $\Sigma N_i = N$. Suppose therefore that $X_{i1},\ldots,X_{iN_i}$ are independent random vectors each distributed according to $p_i(x)$. The proportion of the X_{ij} correctly classified,

$$\hat{p}_{ci} = \frac{1}{N_i}\sum_{j=1}^{N_i} I_i(X_{ij}), \tag{5}$$

is an unbiased estimator of p_{ci}. Hence by (1)

$$\hat{p} = \sum_{i=i}^{M} P_i\hat{p}_{ci} \tag{6}$$

is an unbiased estimator for p_c having variance

$$\sigma^2(\hat{p}) = \sum_{i=1}^{M} \frac{P_i^{\,2}}{N_i}\,(p_{ci} - p_{ci}^{\,2}). \tag{7}$$

How shall we distribute the N observations among the classes? A common choice is to make $N_i = P_iN$, proportional to the prior probabilities. Call the resulting estimator of form (6) $\hat{p}_2$. Note that $\hat{p}_2$ is just the overall proportion of observations correctly classified - that is, $\hat{p}_2$ is the same function of the observations as $\hat{p}_1$, but is obtained from a different sampling design. By (7) we obtain

$$\sigma^2(\hat{p}_2) = \frac{1}{N} \sum_{i=1}^{M} P_i(p_{ci} - p_{ci}^2)$$

$$= \frac{1}{N}(p_c - \sum_{i=1}^{M} P_i p_{ci}^2). \tag{8}$$

Comparing (8) with (4) and applying the fact that for any random variable Z

$$E(Z^2) \geq (EZ)^2 \tag{9}$$

to the random variable taking values p_{ci} with probabilities P_i, we see that $\sigma^2(\hat{p}_2) \leq \sigma^2(\hat{p}_1)$ as expected [2].

The estimator $\hat{p}_2$ would have minimum variance among estimators of the class (6) if only the P_i were known. Since the $p_i(x)$ and hence (in theory) the p_{ci} are known, the optimum choice of N_i is proportional to the product of the prior probability P_i and the within-class standard deviation $\sigma_i = (p_{ci} - p_{ci}^2)^{\frac{1}{2}}$[7]. (This is trivially obtained by applying the Lagrange multiplier method to minimize

$$\sigma^2(\hat{p}) = \sum_{i=1}^{M} \frac{P_i^2 \sigma_i^2}{N_i}$$

subject to the constraint $N_1 + \ldots + N_M = N$). Let $\hat{p}_3$ denote the estimator of form (6) with

$$N_i = \frac{P_i \sigma_i}{\Sigma P_j \sigma_j} N.$$

This optimum estimator has variance

$$\sigma^2(\hat{p}_3) = \frac{1}{N} (\sum_{i=1}^{M} P_i \sigma_i)^2 \tag{10}$$

and $\sigma^2(\hat{p}_3) \leq \sigma^2(\hat{p}_2)$ by another application of (9). The estimator $\hat{p}_3$ is of theoretical interest only, since Monte Carlo estimation of p_c is unnecessary when the p_{ci} can be calculated.

III. POSTERIOR PROBABILITY ESTIMATORS

A different approach to estimation of p_c was discussed in the unclassified samples case by Fukunaga and Kessell [4]. We will extend their idea to classified samples and obtain further variance comparisons. First notice that

$$p_c = \sum_{i=1}^{M} \int P_i I_i(x)\ p_i(x)dx$$

$$= \int \sum_{i=1}^{M} I_i(x)\ p(\omega_i|x) \quad p(x)dx = E[Q(X)] \tag{11}$$

where

$$Q(x) = \sum_{i=1}^{M} I_i(x)\ p(\omega_i|x)$$

is the function which is equal to the posterior probability $p(\omega_i|x)$ of class ω_i when x falls in Γ_i, i = 1,...,M. From (11) it is clear that an unbiased estimator of p_c from unclassified samples $X_1,\ldots,X_N$ is

$$\hat{p}_4 = \frac{1}{N} \sum_{i=1}^{N} Q(X_i).$$

Clearly

$$\sigma^2(\hat{p}_4) = \frac{1}{N}\sigma^2(Q) = \frac{1}{N}(E(Q^2)-p_c^2). \tag{12}$$

Since $0 \leq Q \leq 1$ always, $E(Q^2) \leq E(Q)$ and hence from (12), (4) and (11), $\sigma^2(\hat{p}_4) \leq \sigma^2(\hat{p}_1)$. In fact, in [4] it is shown that for maximum likelihood rules,

$$N[\sigma^2(\hat{p}_1) - \sigma^2(\hat{p}_4)] \geq \frac{1}{M}(1-p_c).$$

This can also be shown by noting that in this case,

$$\sigma^2(p_4) = \frac{1}{N}[E \max_i{}^2\ p(\omega_i|X) - p_c^2],$$

and that in Figure 1,

$$\frac{M+1}{M}\max - \frac{1}{M} \geq \max{}^2,\ \frac{1}{M} \leq \max \leq 1,$$

so that

$$\sigma^2(\hat{p}_4) \leq \frac{1}{N}[(\frac{M+1}{M})p_c - \frac{1}{M} - p_c{}^2],$$

Γ maximum likelihood.

With classified samples $X_{i1},\ldots,X_{iN_i}$ for $i=1,\ldots,M$ and $\Sigma N_i = N$ we can estimate the conditional expected value $E(Q|\omega_i)$ (that is, the expected value of $Q(X)$ when X has density $p_i(x)$) by

$$\frac{1}{N_i}\sum_{j=1}^{N_i} Q(X_{ij}).$$

Since

$$p_c = \sum_{i=1}^{M} P_i \; E(Q|\omega_i) \tag{13}$$

we have a class of unbiased estimators of p_c given by

$$\hat{p} = \sum_{i=1}^{M} P_i \; \left(\frac{1}{N_i}\sum_{j=1}^{N_i} Q(X_{ij})\right) \tag{14}$$

with variances

$$\sigma^2 \; (\hat{p}) = \sum_{i=1}^{M} \frac{P_i^2}{N_i} \sigma^2(Q|\omega_i) \tag{15}$$

where $\sigma^2(Q|\omega_i)$ is the variance of $Q(X)$ when X has density $p_i(x)$.

Special cases are again of interest, the most prominant being the case $N_i = p_i N$. The estimator of form (6) for this allocation of observations is $\hat{p}_5$. Just as $\hat{p}_1$ and $\hat{p}_2$ are the same function computed from different sample designs, so $\hat{p}_5$ is just the mean sample Q and hence equal to $\hat{p}_4$ as a function of the N observations. We obtain from (15) that

$$\begin{aligned}\sigma^2(\hat{p}_5) &= \frac{1}{N}\sum_{i=1}^{M} P_i(E(Q^2|\omega_i) - E(Q|\omega_i)^2) \\ &= \frac{1}{N}\; (E(Q^2) - \sum_{i=1}^{M} P_i \; E(Q|\omega_i)^2).\end{aligned} \tag{16}$$

Applying (9) to the second terms of (16) and (12) in the light of (13) shows that $\sigma^2(\hat{p}_5) \leq \sigma^2(\hat{p}_4)$.

The optimal choice of N_i is proportional to $P_i \; \sigma(Q|\omega_i)$. The corresponding estimator of form (14) has minimum variance in that class. Denoting this estimator by $\hat{p}_6$,

$$\sigma^2(\hat{p}_6) = \frac{1}{N}\; (\sum_{i=1}^{M} P_i\sigma(Q|\omega_i))^2$$

and $\sigma^2(\hat{p}_6) \le \sigma^2(\hat{p}_5)$ by (9).

IV. COMPARISON OF VARIANCES

If we use $\hat{p}_i << \hat{p}_j$ to mean that $\hat{p}_j$ dominates $\hat{p}_i$ in the sense of having variance no greater than the variance of $\hat{p}_i$ for all choices of M, P_i, Γ_i and $p_i(x)$, then Fig. 2 summarizes our results to this point.

It is natural to hope that $\hat{p}_5 >> \hat{p}_2$. This is false, for we now give an example to show that <u>no uniform dominance exists between $\hat{p}_5$ and $\hat{p}_2$</u>.

Consider the two-class problem with $\omega_1 \neq \omega_2$ both real numbers $0 \le \omega_i \le 1$, X a Bernoulli variable with density

$$p_i(x) = \omega_i \qquad x = 0$$
$$= 1-\omega_i \qquad x = 1,$$

and $P_1 = 1-P_2 = P$. There are only four possible classification rules. These can be described by Γ_1, the set of observations which will lead to classification as ω_1, as follows,

δ_1: $\Gamma_1 = \{0\}$

δ_2: $\Gamma_1 = \{1\}$

δ_3: $\Gamma_1 = \{0,1\}$

δ_4: $\Gamma_1 = \phi$.

Consider first the rule δ_3 which always classifies X as ω_1. In this case $p_c = P$, $\hat{p}_{c1} = 1$, $\hat{p}_{c2} = 0$ and hence $\hat{p}_2 \equiv P$ by (6). Thus $\sigma^2(\hat{p}_2) = 0$, and indeed any estimator of form (6) has variance 0. Turning to $\hat{p}_5$, note that

$$Q(0) = p(\omega_1 | X = 0) = \frac{\omega_1 P}{\omega_1 P+\omega_2(1-P)}$$

$$Q(1) = p(\omega_1 | X = 1) = \frac{(1-\omega_1)P}{(1-\omega_1)P+(1-\omega_2)(1-P)} .$$

Hence Q(x) is not constant and we see from (16) that therefore $\sigma^2(\hat{p}_5) > 0 = \sigma^2(\hat{p}_2)$. Thus $\hat{p}_5$ does not dominate $\hat{p}_2$. Notice in particular that if P is large enough, specifically, if

$$\frac{P}{1-P} > \max\{\frac{\omega_2}{\omega_1}, \frac{1-\omega_2}{1-\omega_1}\},$$

then δ_3 is the Bayes classification rule in this example. So $\hat{p}_5$ is not even always preferable to $\hat{p}_2$ when the optimum classification rule is used.

To show that $\hat{p}_2$ does not dominate $\hat{p}_5$, consider the classification rule δ_1 and the case $P = 1/2$, $\omega_1 = 1/2$, $\omega_2 = 0$. Computation shows that in this case $N\,\sigma^2(\hat{p}_2) = 1/8$ while $N\,\sigma^2(\hat{p}_5) = 1/72$. A similar absence of uniform dominance applies to $\hat{p}_3$ and $\hat{p}_6$, as can be demonstrated by the same pair of examples.

V. EMPIRICAL RESULTS

The lack of a definitive relationship between $\hat{p}_2$, the selective sampling statistic, and the posterior estimators leads us to the consideration of test cases. Since $\hat{p}_1$, $\hat{p}_2$, $\hat{p}_4$, and $\hat{p}_5$ are likely to be of most interest in simulation experiments, we will consider only these four statistics.

Consider the problem of estimating p_c for a multiple-hypothesis testing problem involving three equally-likely univariate-normal classes with unit variance and means 0.0, 0.5, and 3.25. Initially, assume that a decision rule which is optimal (in the sense of error) is desired. The corresponding rule and the densities as well as the posterior probabilities are depicted in Figure 3. The decision boundaries are 0.25 and 1.875, and p_c is 0.6761.

In the experiment to determine the relative effectiveness of the four statistics, the sample variance for each was computed for 500 trials, using 30, 60,..., 570, and 600 computed generated pseudo-random numbers (each set included the previous). Two different sets of numbers were used, one for $\hat{p}_1$ and $\hat{p}_4$, and another for $\hat{p}_2$ and $\hat{p}_5$. In the former case, sampling from the mixture density was simulated by the additional step of choosing $p_i(x)$ according to the priors P_i, again, by using pseudo-random numbers. In the latter case, an equal number of samples were generated for each $p_i(x)$.

The results for this optimum rule are given in Figure 4. Since $\sigma(\hat{p}_1)$ and $\sigma(\hat{p}_2)$ can be computed exactly, their values are included (dashed lines) for comparison purposes. In this case, we see that both posterior statistics perform significantly better than the counting statistics, with the selective posterior statistic somewhat better than the posterior statistic employing unclassified samples.

In another experiment, the same procedure was repeated, using a new set of pseudo-random numbers for a suboptimal decision rule. The same three densities were used, but the decision boundaries were changed to -0.5 and 2.5. In this case, p_c = .6335. The results are given in Figure 5. Sample variance for each statistic increased slightly, but the observations made in the first experiment still apply.

VI. SUMMARY

Variance relationships among several estimators of probability of correct recognition p_c, employing both classified and unclassified samples, were discussed. A statistic for p_c based on a stratified (selective) sampling design and posterior probabilities was introduced. Experimental evidence of the utility of this statistic was presented. A possible drawback in the use of estimators using posterior probabilities is the requirement that class density functions must be known. However, the use of density estimation methods, and the fact that in many Monte-Carlo studies, densities are known, tend to point out the usefulness of these statistics.

BIBLIOGRAPHY

1. K. Fukunaga, Introduction to Statistical Pattern Recognition. New York: Academic Press, 1972.

2. W. Highleyman, "The design and analysis of pattern recognition experiments", Bell Syst. Tech. J., vol. 41, pp. 723-744, 1963.

3. H. Cramér, The Elements of Probability Theory and Some of Its Applications. New York: Wiley, 1955.

4. K. Fukunaga and D. Kessell, "Application of optimum error reject functions", IEEE Trans. Inform. Theory, vol. IT-18, pp. 814-817, Nov. 1972.

5. C. Chow, "On optimum recognition error and reject tradeoff", IEEE Trans. Inform. Theory, vol. IT-16, pp. 41-46, Jan. 1970.

6. K. Fukunaga and D. Kessell, "Non-parametric Bayes error estimation using unclassified samples", IEEE Trans. Inform. Theory, vol. IT-19, pp. 434-440, July 1973.

7. J. Neyman, "On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection", J. Royal Statist. Soc., vol. 97, pp. 558-625, 1934.

FIGURE CAPTIONS

Figure 1. Upper bound on $\max_i^2 p(\omega_i|x)$

Figure 2. Dominance relations among estimators of p_c

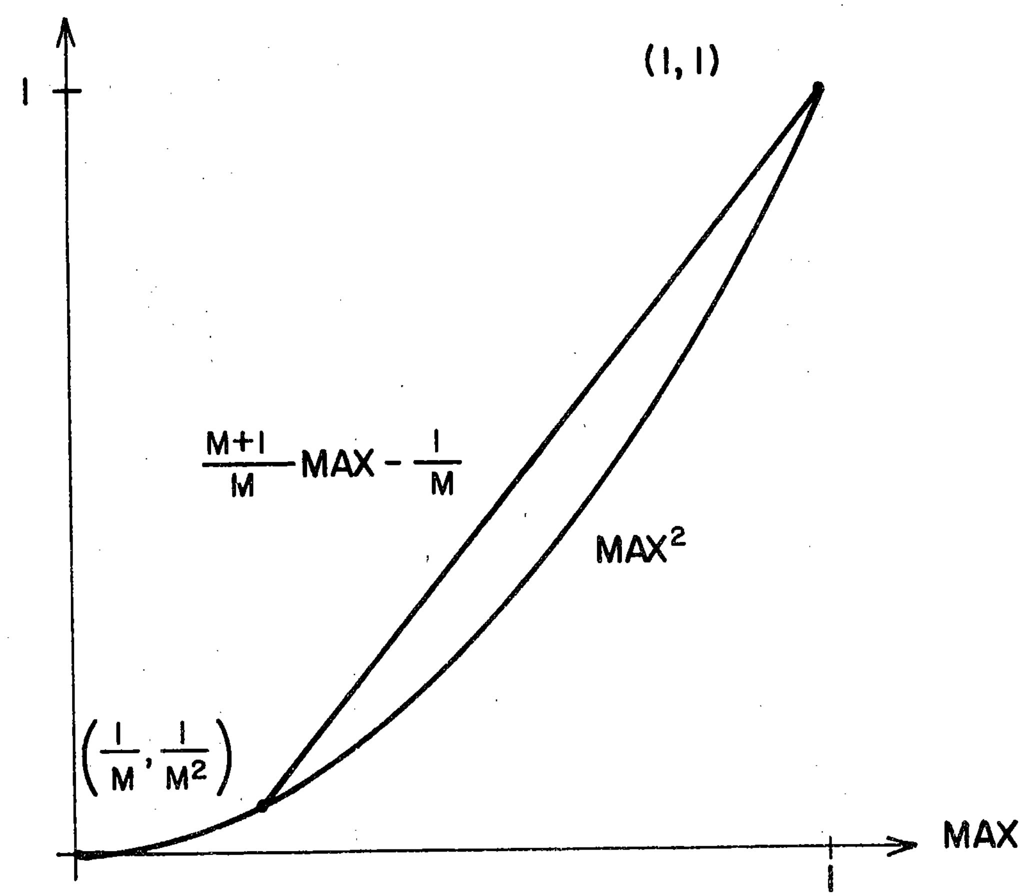
Figure 3. Optimum rule for three normal classes

Top-mixture density and decision rule

Bottom-maximum posterior probabilities

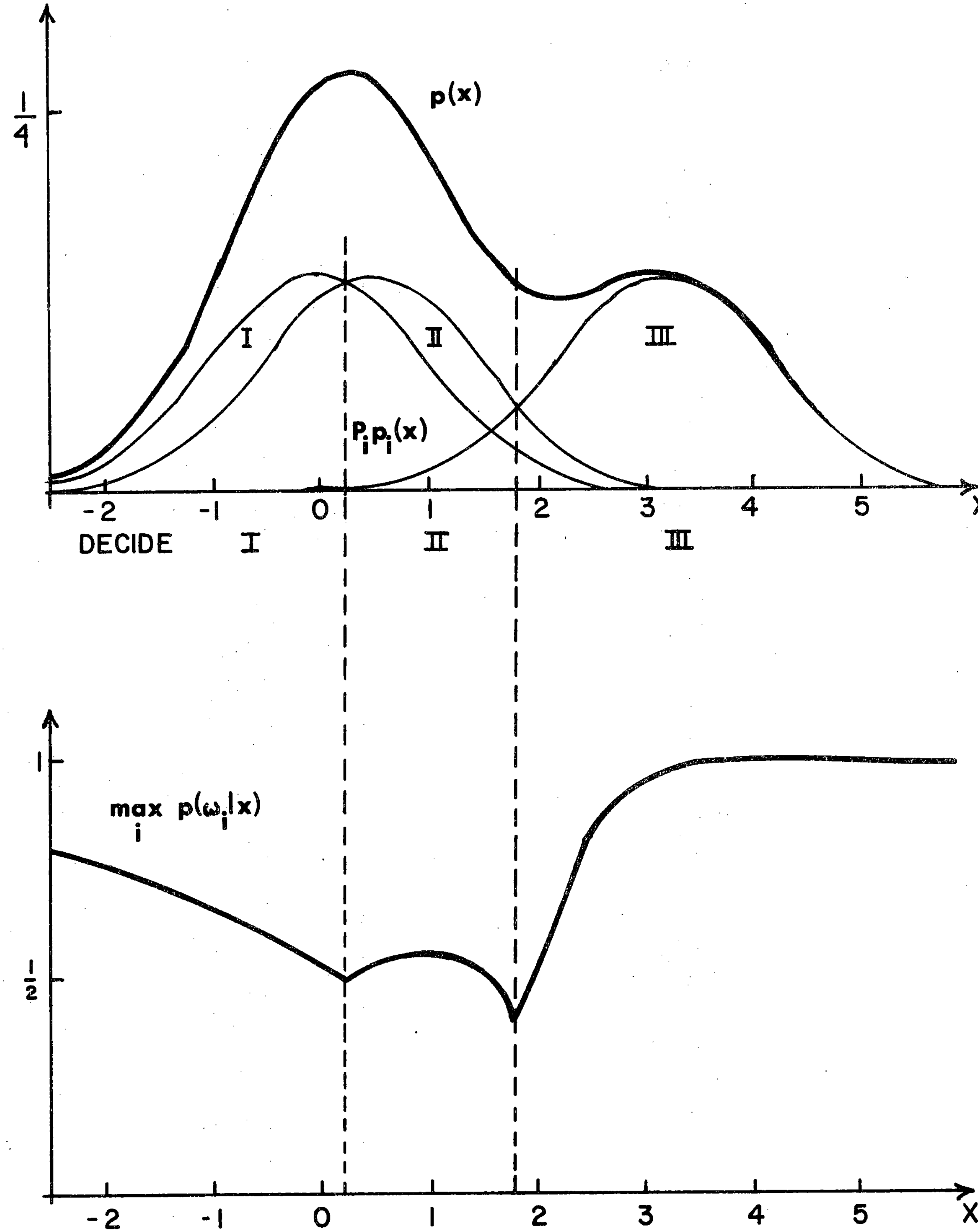
Figure 4. Sample standard deviation of four estimators of p_c for an optimal rule - p_c=.6761

Figure 5. Sample standard deviation of four estimators of p_c for a sub-optimal rule - p_c=.6335

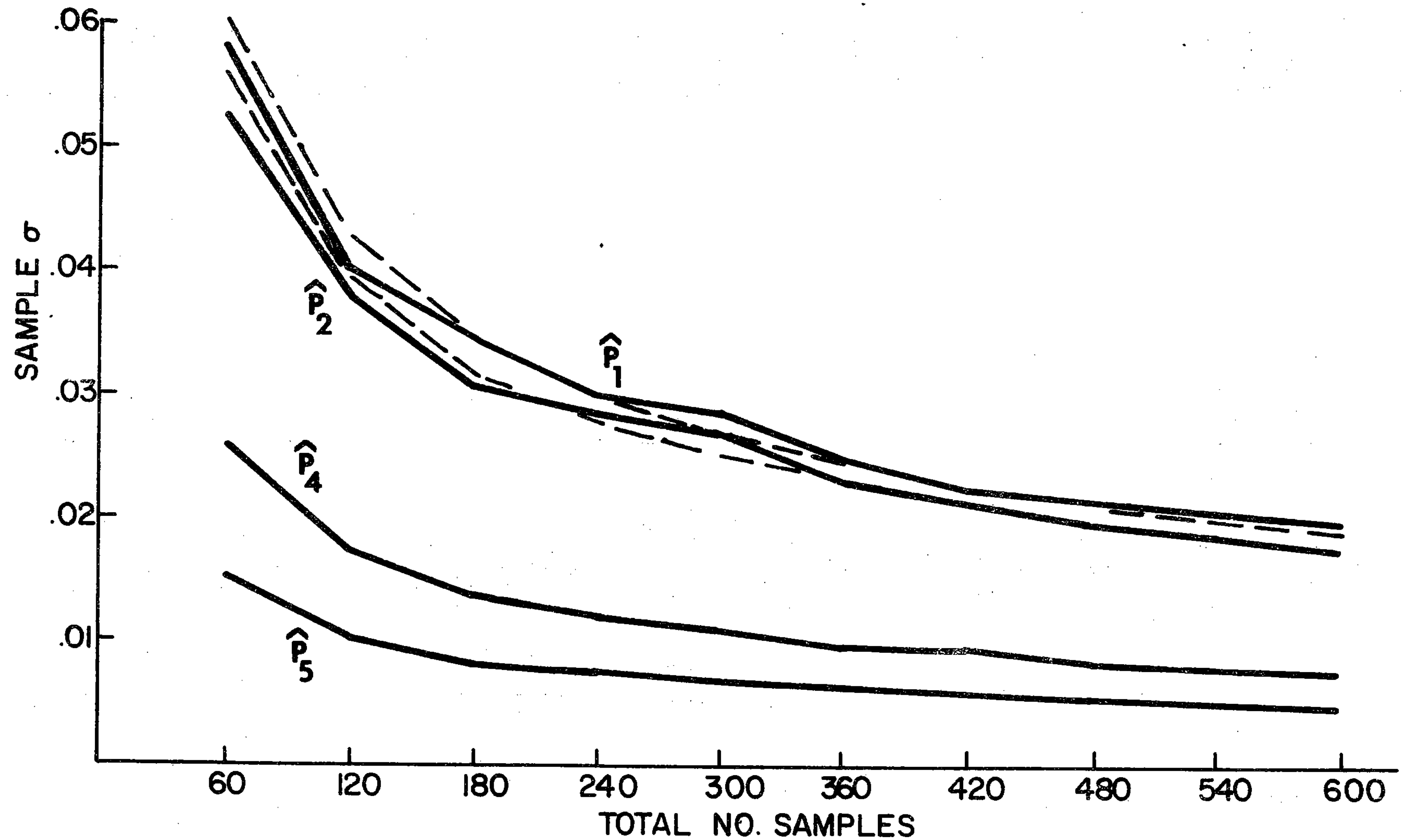


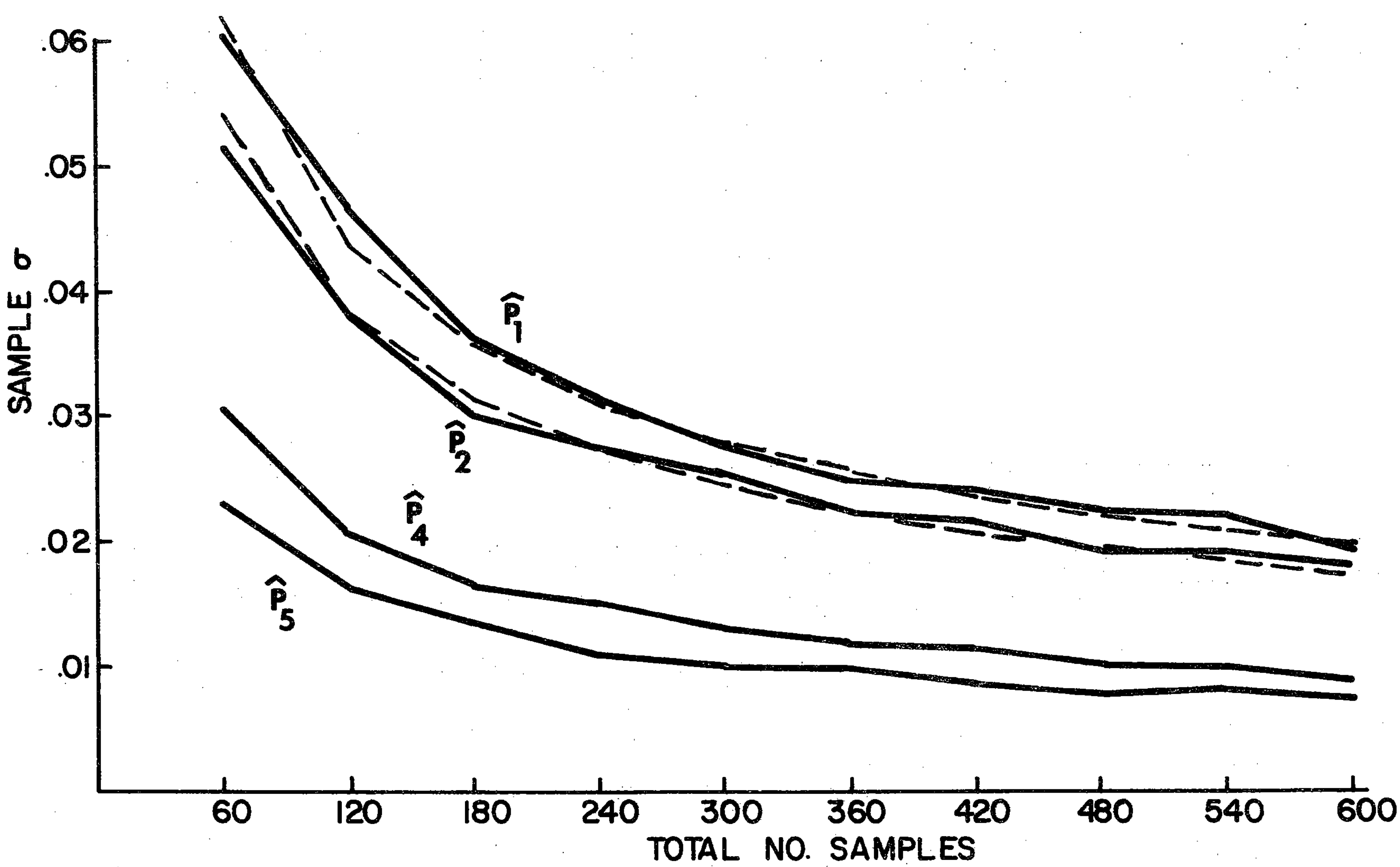
(1, 1)
1
$\frac{M+1}{M}MAX - \frac{1}{M}$
MAX^2
$\left(\frac{1}{M}, \frac{1}{M^2}\right)$
MAX
1

	Unclassified Samples		Classified $N_i=P_iN$		Classified, Optimal N_i
Count Estimators	$\hat{p}_1$	$<<$	$\hat{p}_2$	$<<$	$\hat{p}_3$
	$\overset{\wedge}{\wedge}$				
Posterior Estimators	$\hat{p}_4$	$<<$	$\hat{p}_5$	$<<$	$\hat{p}_6$



p(x)
1/4
I
II
III
$P_i p_i(x)$
-2
-1
0
1
2
3
4
5
X
DECIDE
I
II
III
1
$\max_i p(\omega_i|x)$
1/2
-2
-1
0
1
2
3
4
5
X


SAMPLE σ
.06
.05
.04
.03
.02
.01
$\hat{P}_1$
$\hat{P}_2$
$\hat{P}_4$
$\hat{P}_5$
60
120
180
240
300
360
420
480
540
600
TOTAL NO. SAMPLES



SAMPLE σ
.06
.05
.04
.03
.02
.01
60
120
180
240
300
360
420
480
540
600
TOTAL NO. SAMPLES
$\hat{P}_1$
$\hat{P}_2$
$\hat{P}_4$
$\hat{P}_5$

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
Mimeograph Series #401

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Variance Comparisons for unbiased Estimators of Probabilities of Correct Classifications

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER
Mimeograph Series #401

7. AUTHOR(s)
David S. Moore, Stephen J. Whitsitt and David A. Landgrebe

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Purdue University, W. Lafayette, IN 47907

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
6269

11. CONTROLLING OFFICE NAME AND ADDRESS
Air Force Office of Scientific Research
Washington, D.C.

12. REPORT DATE
January 1975

13. NUMBER OF PAGES
17

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Distribution of this document is unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
classification rules
probability of correct classification
unbiased estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Variance relationships among certain count estimators and posterior-probability estimators of correct recognition are investigated. A statistic using posterior probabilities is presented for use in stratified sampling designs. A test case involving three normal classes is examined.