Comp. & Maths. with Appls. Vol. 12A, No. 2, pp. 289–299, 1986
Printed in Great Britain.

0097-4943/86 $3.00 + .00

ESTIMATING CLASS SIZES BY ADJUSTING FALLIBLE CLASSIFIER RESULTS

D. J. HAND
Biometrics Unit, Institute of Psychiatry, De Crespigny Park, London SE5 8AF, England

Abstract—In discriminant analysis, class sizes are usually estimated by the proportion of a random sample which falls into each class. The accuracy of this method is limited by the low practicability of applying the ideal classification rule to large numbers of objects. How this basic estimator may be improved by taking advantage of the relative feasibility of applying the derived classification rule to large numbers of objects is discussed. Two methods are outlined and comparisons between them are made, resulting in a recommendation that one of the methods should be preferred.

1. INTRODUCTION

This paper addresses the problem of estimating the sizes of the classes comprising a population. We adopt the usual discriminant analysis and statistical pattern recognition formulation[4]: we suppose there exists some infallible classification rule (commonly called, in pattern recognition terminology, a "teacher", but which we shall simply refer to as the perfect classification rule), that this has been applied to a sample of objects, and that from this correctly classified sample a new classification rule has been derived. Typically this new rule is imperfect—it misclassifies some of the objects—but it is also more practicable to apply than the perfect rule. "More practicable" here may mean less expensive, but it can also have other meanings as we illustrate in the next section. Whatever meaning is used, we intend it to have the implication that it is feasible to apply the imperfect classifier to a large sample, but it is not feasible to apply the perfect classifier to a large sample.

In common with much other discriminant analysis work we also assume we have a test set. This is a set of objects, independent of the set used to formulate the decision rule, sampled (using a simple random sampling scheme) from the same population and which has also been classified by both rules. This set provides an independent assessment of the performance of the imperfect classification rule. It also provides a sample of objects cross-classified by the perfect and imperfect rules.

Given this situation, our aim is to apply either the perfect rule or the imperfect rule, or both, to yield an estimate of the sizes of the classes in the population being classified. It is worth noting that this population could but need not be the original one from which the design sample was taken. The test set, however, must be sampled from the population being studied.

There are two obvious elementary approaches:

(i) One can simply examine the results of the perfect classifier applied to the test set. This ignores the imperfect classifier.
(ii) One can apply the imperfect classifier to a large sample (feasible in view of the discussion above), a sample which might include the test set. Typically this large sample would be the objects to which the classifier is applied when it becomes operational. Apart from the role of the perfect classifier in designing the imperfect rule (and in classifying the test set) this method ignores the perfect classifier.

Since methods (i) and (ii) each ignore part of the available information one is naturally led to ask whether they can be combined to yield a more effective method. This paper discusses two such methods and explores their relative merits.

After introducing the problem in its general form, for reasons of expository convenience as well as because this is the single most common special case, much of the discussion below will be restricted to the case of two classes. In particular this means that we can concern ourselves with estimating just the proportion of objects in one class, the other estimator having symmetric properties and being obtained by subtraction.

2. AREAS OF APPLICATION

The problem of estimating class sizes when both perfect and imperfect classifiers are available, as described in Sec. 1, arises in a great number of domains. Some examples are as follows:

(i) Disease prevalence estimation is a very common application. Typically the "perfect" classification is a full medical examination, perhaps involving lengthy and expensive tests (CAT scanning, NMR scanning, histological examination, bacteriological examination, etc.) Cost alone could prevent this from being used on a large sample, but another reason could be the length of time the tests involve. In a psychiatric environment, the "perfect" classification is typically an interview with a trained and standardised psychiatrist. The imperfect rule in this kind of situation could be based on the results of a few quick or cheap tests or could be a score on a questionnaire. This latter is common in psychiatry and some other areas of medicine.

(ii) Estimating the proportion of faulty items in quality control is another example. The aim is to decide whether the production system is functioning within acceptable bounds, or whether it is producing too many faulty products. Identification of individual faulty items is not the objective. The perfect classifier might involve extremely complex and time-consuming sets of tests, whereas the imperfect classifier might utilise only a few relatively straightforward tests.

(iii) Of course, any situation in which the perfect classifier destroys or damages the objects can gain from the kind of approaches discussed in this paper. Examples of this occur in quality control and in biological studies (e.g. exploring the effect of drugs on metabolism. This frequently involves killing animals). The imperfect classifier will produce its classification with comparatively mild consequences.

(iv) Apart from epidemiological examples, as in (i) above, many social survey questions can benefit from this kind of approach. For example, one might be trying to correct bias in stated voting intention (the imperfect classifier) through the use of a complex model of attitude which involves extensive questioning of a respondent (the perfect classifier).

(v) Many predictive situations fall into this class. For example the perfect classification is death or survival after 5 years, and the imperfect classifier is a questionnaire given now.

3. THE ESTIMATORS

The data for the estimators consist of two parts. The first is a random sample size m from the population under study. This sample has been classified only by the imperfect classifier, yielding m_i classified into class i. $\mathbf{M}$ is the vector with ith element m_i. The second part of the data consists of a random sample of size n (the test set), to which both classifiers have been applied. n_{ij} of the elements are classified simultaneously as class i by the imperfect classifier and class j by the perfect classifier. $\mathbf{N}$ is the matrix with ijth element n_{ij}. Generally, as implied in Section 1, $m \gg n$, so that we shall assume m infinite in much of what follows.

We let $\boldsymbol{\pi}$ be the vector of true class sizes, and $\hat{\boldsymbol{\pi}}$ be an estimator.

Then in terms of this notation the two elementary estimators introduced in Section 1 are as follows.

Estimator solely using perfect classifier

This is the first case of Sec. 1 and is the estimator based only on the test set simple random sample of size n. Then

$$\hat{\boldsymbol{\pi}} = \mathbf{N}'\mathbf{1}/n \quad (\mathbf{1} \text{ is a vector of 1's}).$$

Estimator solely using imperfect classifier

The second case of Sec. 1.

$$\hat{\boldsymbol{\pi}} = \mathbf{M}/m.$$

Each of the above estimators only makes use of part of the available information. We now introduce alternatives which make better use of the available data. In the asymptotic case of m infinitely large, we shall put $\hat{\boldsymbol{\pi}} = \mathbf{M}/m = \mathbf{P}$.

Estimator using probabilities conditional on imperfect classifications

If $\mathbf{A}$ is the matrix whose ijth element is the probability of having a true (perfect) classification i when the imperfect classifier assigns to class j, then

$$\boldsymbol{\pi} = \mathbf{A}'\mathbf{P}.$$

The estimator based on this identity is

$$\hat{\boldsymbol{\pi}}^A = \mathbf{N}' \operatorname{diag}(\mathbf{N1})^{-1}\mathbf{P},$$

where $\operatorname{diag}(\mathbf{X})$ is the diagonal matrix with the vector $\mathbf{X}$ down its leading diagonal.

Estimator using probabilities conditional on perfect classifications

If $\mathbf{B}$ is the matrix whose ijth element is the probability that the imperfect classifier assigns an object to class i, given that the perfect classifier has assigned it to class j, then

$$\boldsymbol{\pi} = \mathbf{B}^{-1}\mathbf{P}.$$

The estimator derived from this is

$$\hat{\boldsymbol{\pi}}^B = \operatorname{diag}(\mathbf{N}'\mathbf{1})\mathbf{N}^{-1}\mathbf{P}.$$

The reader will note from the definitions of $\mathbf{A}$ and $\mathbf{B}$ that cross-classifications of simple random sample test sets are not in fact necessary for $\hat{\boldsymbol{\pi}}^A$ and $\hat{\boldsymbol{\pi}}^B$. For example, in the first case all that is necessary is that the probabilities of perfect results conditional on imperfect results can be estimated. This could be estimated from a wide variety of sampling schemes. For example, we could force each row of $\mathbf{N}$ to sum to the same total, N, say. That is, the estimated conditional probabilities within any given row are obtained by using the perfect classifier to classify N objects known to have imperfect classifications assigning them to this row. Or one might choose the row totals proportional to a prior estimate of within row variation, or proportional to the observed m_i values. Similar comments apply to the column conditional probabilities in $\mathbf{B}$. Such possibilities have various merits and are worth consideration. However, the fact is that in discriminant analysis and statistical pattern recognition work, in general, a simple randomly sampled test set seems to be by far the most common case.

Further discussion of these points may be found in [1,3,5,7–10]. Much of this work describes the two-class case. In particular, the reader should note that any results on using stratified sampling or double sampling schemes to estimate means of interval scale variables can be applied in this case because a two-class (binary) variable can be regarded as an interval variable. Particularly relevant here are the results in [2].

In the next section we consider the properties of the above estimators for the two-class case in some detail. As noted in Sec. 1, in the two-class case we can simply concern ourselves with estimating the size of one class, since the other's size is complementary. In what follows we give results for Class 2.

It is now useful to abandon the earlier general matrix notation and restate the two-class expressions algebraically. The fact that we are only dealing with two classes permits certain notational simplifications. Thus, in what follows, we let π represent the true proportion of objects which are Class 2 (i.e. this is the value which would be obtained were the perfect classifier to be applied to all objects).

Other notation is given in Fig. 1. In particular

$$\theta_1 = P(\text{perfect classification} = 2 | \text{imperfect classification} = 1),$$

$$\theta_2 = P(\text{perfect classification} = 1 | \text{imperfect classification} = 2),$$

$$\phi_1 = P(\text{imperfect classification} = 2 | \text{perfect classification} = 1),$$

$$\phi_2 = P(\text{imperfect classification} = 1 | \text{perfect classification} = 2),$$

From these, using the data in the cross-classification matrix **N**, we get

$$\hat{\theta}_1 = n_{12}/(n_{11} + n_{12}),$$

$$\hat{\theta}_2 = n_{21}/(n_{21} + n_{22}),$$

$$\hat{\phi}_1 = n_{21}/(n_{11} + n_{21}),$$

$$\hat{\phi}_2 = n_{12}/(n_{12} + n_{22}).$$

Using this notation, the two adjusted estimators become

$$\hat{\pi}^A = \hat{P}(1 - \hat{\theta}_2) + (1 - \hat{P})\hat{\theta}_1,$$

$$\hat{\pi}^B = (\hat{P} - \hat{\phi}_1)/(1 - \hat{\phi}_1 - \hat{\phi}_2),$$

with $\hat{P} = m_2/m$.

One could explore the distributions of $\hat{\pi}^A$ and $\hat{\pi}^B$ by transforming the multinomial variables n_{11}, n_{12}, n_{21} and n_{22} to two new sets containing, respectively, the functions $\hat{\pi}^A = \hat{\pi}^A(\mathbf{N})$ and $\hat{\pi}^B = \hat{\pi}^B(\mathbf{N})$, and then integrating out the unwanted functions. It seems, however, that this will lead to a complicated expression. Rather than this, we have chosen to pursue an analytic approximate large-sample approach and a computer enumeration small-sample approach.

4. PROPERTIES OF THE ESTIMATORS

The simple binomial estimator based solely on the test set of size n has the merit of being simple. It is also unbiased. However, since n is typically not large, its variance of $\pi(1 - \pi)/n$ could be excessive.

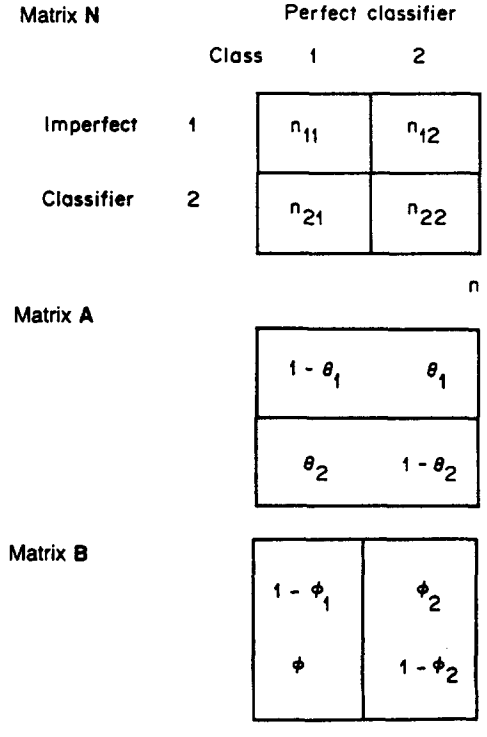


Fig. 1. Notation for the two-class case.

Complementary to this, by virtue of the large size of M, the binomial estimate based solely on the imperfect classifier has small (and we take it to be zero) variance. However, it is typically biased.

$\hat{\pi}^A$ is unbiased, and Tenenbein[7] gives the large sample variance. Comparisons between $\hat{\pi}^A$ and the two elementary estimators may be found in the references given in Sec. 3. In this paper we concentrate chiefly on the comparative properties of $\hat{\pi}^A$ and $\hat{\pi}^B$.

$\hat{\pi}^B$ is more interesting. This estimator can behave strangely. If **B** is singular, $\hat{\pi}^B$ is not defined. In the two-class case this occurs when $\hat{\phi}_1 + \hat{\phi}_2 = 1$, and it is generally the case if any of the imperfect classifier conditional distributions are linear functions of the others. In fact, things are rather worse than this because such linear dependence can occur in the sample matrix **N** even if it does not occur in **B**. This means that although π^B may be well defined, $\hat{\pi}^B$ is not.

This sample dependence can occur whenever **B** is such that it cannot be arranged into diagonal form by permuting rows and columns. In such cases there is always a nonzero (though perhaps very small) probability of **N** being singular. The implication of this is that there is a nonzero probability that $\hat{\pi}^B$ will be infinite (via $\hat{\phi}_1 + \hat{\phi}_2 = 1$ in the two-class case). Hence, whenever **B** cannot be arranged into diagonal form by permuting rows and columns, $\hat{\pi}^B$ has infinite variance. If **B** could be so arranged then, by simple relabelling of the imperfect classifier's classes, we would obtain a perfect classifier. Thus straightforward variance is not a useful descriptor of the distribution of $\hat{\pi}^B$.

White and Castleman[10] use the $\hat{\pi}^B$ form and side-step the problem by assuming **N** fixed and error-free. (For this case they give an expression for asymptotic variance.) However, in our experience it is the variation in **N** which is typically the dominant factor, rather than that of **M**. This arises because of the relative ease (and inexpensiveness) of applying the imperfect classifier compared with the perfect one.

$\hat{\pi}^B$ infinite is not the only situation in which it produces unacceptable values. More generally than this, anything outside the bounds 0 and 1 is useless. This means we must have

$$0 \leq \frac{\hat{P} - \hat{\phi}_1}{1 - \hat{\phi}_1 - \hat{\phi}_2} \leq 1,$$

which is only true if $\hat{P}$ can be expressed as a linear combination of $\hat{\phi}_1$ and $(1 - \hat{\phi}_2)$:

$$\hat{P} = (1 - \alpha)\hat{\phi}_1 + \alpha(1 - \hat{\phi}_2)$$

with $0 \leq \alpha \leq 1$. [The reader is reminded that $(1 - \hat{\phi}_1 - \hat{\phi}_2)$ can be negative.] The symmetry between the role of the pair (π, P) in $\hat{\pi}^A$ and $(P, \pi = \alpha)$ in $\hat{\pi}^B$ is evident here. Since the samples for $\hat{P}$ and the $\hat{\phi}_i$ are independent, there is no guarantee that this condition will be satisfied.

This discussion suggests that $\hat{\pi}^B$ is a worse choice than $\hat{\pi}^A$. However, it might be the case that there are other criteria for comparison, judged by which $\hat{\pi}^B$ is superior. In the next section we explore this suggestion. Yet a further possibility is that some pragmatic modification of $\hat{\pi}^B$ might yield an estimator exhibiting advantages. This is explored in Sec. 6.

5. LARGE-n COMPARISON

When n is large, for reasonable values of π, ϕ_1 and ϕ_2, the probability that $\hat{P}$ falls outside $\hat{\phi}_1$ and $(1 - \hat{\phi}_2)$ can be vanishingly small. This suggests that variance might be an inappropriate criterion in which to base a comparison. (No matter how small is the probability that $\hat{\pi}^B = \infty$, if it is nonzero then the variance is infinite.) As an alternative, we examined the rate of change of $\hat{\pi}^A$ and $\hat{\pi}^B$ with ϕ_1 and ϕ_2 in the neighbourhood of the true ϕ_i. This shows how rapidly $\hat{\pi}^A$ and $\hat{\pi}^B$ deviate from the true values. (We used the ϕ_i, rather than the θ_i, because the former have a more obvious intuitive interpretation.)

The gradient vectors

$$\frac{\partial \pi^A}{\partial \phi_1}, \quad \frac{\partial \pi^A}{\partial \phi_2}$$

and

$$\frac{\partial \pi^B}{\partial \phi_1}, \quad \frac{\partial \pi^B}{\partial \phi_2}$$

are given by (π^A) using

$$\theta_1 = \phi_2\pi/[(1 - \phi_1)(1 - \pi) + \phi_2\pi],$$
$$\theta_2 = \phi_1(1 - \pi)/[\phi_1(1 - \pi) + (1 - \phi_2)\pi],$$

and

$$\frac{\partial \pi^A}{\partial \phi_1} = \frac{\partial \pi^A}{\partial \theta_1} \cdot \frac{\partial \theta_1}{\partial \phi_1} + \frac{\partial \pi^A}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial \phi_1},$$
$$\frac{\partial \pi^A}{\partial \phi_2} = \frac{\partial \pi^A}{\partial \theta_1} \cdot \frac{\partial \theta_1}{\partial \phi_2} + \frac{\partial \pi^A}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial \phi_2},$$

with

$$\frac{\partial \pi^A}{\partial \theta_1} = 1 - P, \quad \frac{\partial \pi^A}{\partial \theta_2} = P,$$
$$\frac{\partial \theta_1}{\partial \phi_1} = \frac{\phi_2\pi(1 - \pi)}{[(1 - \phi_1)(1 - \pi) + \phi_2\pi]^2},$$
$$\frac{\partial \theta_2}{\partial \phi_1} = \frac{(1 - \phi_2)\pi(1 - \pi)}{[\phi_1(1 - \pi) + (1 - \phi_2)\pi]^2},$$
$$\frac{\partial \theta_1}{\partial \phi_2} = \frac{(1 - \phi_1)\pi(1 - \pi)}{[(1 - \phi_1)(1 - \pi) + \phi_2\pi]^2},$$
$$\frac{\partial \theta_2}{\partial \phi_2} = \frac{\phi_1(1 - \pi)\pi}{[\phi_1(1 - \pi) + (1 - \phi_2)\pi]^2},$$

and π^B:

$$\frac{\partial \pi^B}{\partial \phi_1} = \frac{(1 - \phi_2 - P)}{[1 - \phi_1 - \phi_2]^2},$$
$$\frac{\partial \pi^B}{\partial \phi_2} = \frac{P - \phi_1}{[1 - \phi_1 - \phi_2]^2},$$

where $P = \phi_1 + \pi(1 - \phi_1 - \phi_2)$.

Now it is clear from these expressions that the direction of maximum change of $\hat{\pi}^A$ (given by its gradient vector with respect to ϕ_1 and ϕ_2) is not the same as that of $\hat{\pi}^B$. This means that there are some directions of $\boldsymbol{\phi}$ perturbation for which $\hat{\pi}^A$ deviates from the true value more rapidly than does $\hat{\pi}^B$. That is, for some classes of ϕ_i change, $\hat{\pi}^B$ is better. In practice, of course, it would be impossible to identify these. We therefore considered the change induced in $\hat{\pi}^A$ by a small step in the direction of its maximum change compared with the change induced in $\hat{\pi}^B$ by a small step of the same length in the direction of its maximum change. Computer search failed to locate any (π, ϕ_1, ϕ_2) combination for which $\hat{\pi}^A$ had a larger maximum change than $\hat{\pi}^B$.

To illustrate this, consider the special case of $\phi_1 \simeq \phi_2 \simeq \pi \simeq \delta \ll 1$, such that δ^2 can be neglected. Then from the above derivatives we find that the change in $\hat{\pi}^A$ due to a small step

ϵ in the direction of maximum change is $(\frac{1}{2} + \frac{3}{2}\delta)\epsilon$. For $\hat{\pi}^B$, however, the change due to a small step ϵ in the direction of its maximum change is ϵ. Thus $\hat{\pi}^A$ appears to degrade less due to sampling fluctuations from the true ϕ_i. That is, $\hat{\pi}^A$ is less sensitive to small sampling fluctuations than is $\hat{\pi}^B$.

6. SMALL-n COMPARISON

Small values for n mean that the probability of $\hat{\pi}^B$ taking an unacceptable value may not be near zero. However, such cases can be recognised if and when they occur. Thus it is worth exploring the properties of the restricted $\hat{\pi}^B$, which is only defined when it lies in the permissible region. It might be the case that when one modifies the estimator in this way it leads to a function having properties more desirable than those of $\hat{\pi}^A$, for those values for which it is defined. Of course, this also applies if n is large.

We used computer enumeration to explore this. For given values of π, ϕ_1, ϕ_2 and n, the computer worked through all assignments of the n points to the four cells of the 2×2 (imperfect by perfect) cross-classification. There are $(n + 3)(n + 2)(n + 1)/6$ such assignments. [This can easily be seen by imagining the n objects arrayed in a line. One must then place three partitions between them—to divide them into four cells. Since the objects are indistinguishable and the partitions are indistinguishable we have

$$\frac{(n + 3)!}{n!3!}$$

distinct assignments.]

Again for pragmatic reasons, partitions leading to zero marginals were skipped. This means, of course, that the resulting estimators will typically (except in rare symmetric cases) be biased. This is evident in Table 1, below, especially in the small-π cases. For the others the program calculated the probability of obtaining the distribution of n_{ij}:

$$\frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\pi^{n_{12}+n_{22}}(1 - \pi)^{n_{11}+n_{21}}\phi_1^{n_{21}}(1 - \phi_1)^{n_{11}}\phi_2^{n_{12}}(1 - \phi_2)^{n_{22}}$$

For this N, the estimates $\hat{\pi}^A$ and $\hat{\pi}^B$ were obtained.

The estimates and the probabilities were then combined to yield overall values for variance and mean-square error: the two criteria used in this part of the study. (Note in calculating expectations, rescaling is needed to allow for the skipping of the zero marginal cases.) Table 1 is illustrative of the results obtained. [In fact, the results in Table 1 are based on a further pragmatic modification. If $|1 - \hat{\phi}_1 - \hat{\phi}_2| < 10^{-5}$ then the estimate was rejected (as either singular or, at the least, unreliable). We are bending over backwards to give $\hat{\pi}^B$ the best conditions for demonstrating its worth.] It is clear that even with this modified version of $\hat{\pi}^B$, $\hat{\pi}^A$ is the superior estimator.

Insight can be gained into the shapes of the distributions by studying plots. We are dealing with discrete distributions, but the large number of points at which the probability is nonzero [calculated from $(n + 3)(n + 2)(n + 1)/6$] means that we must resort to a summarising display. Since each nonzero point has an associated probability generated by the computer enumeration, an obvious first choice for the display would be a histogram. Unfortunately such an approach seems very sensitive to the choice of position for the cell boundaries. Unlike standard applications of histograms, in which each nonzero sample point contributes a probability of $1/n$, in the present case each point contributes a different probability—and in some cases it can be substantial. In such circumstances, computational rounding error can be critical.

We therefore used a modified kernel density estimate to produce the plots of Figs. 2–5[5]. Its form was

$$\hat{f}(\hat{\pi}) = \sum_j \text{prob}(\rho_j)K(\hat{\pi}, \rho_j),$$

Table 1. Comparison of $\hat{\pi}^1$ and $\hat{\pi}^B$ using modified $\hat{\pi}^B$ estimator which discounts points falling outside [0, 1]

π	ϕ_1	ϕ_2	n	$\frac{\text{MSE}(A)}{\text{MSE}(B)}$	$\frac{\text{Var}(A)}{\text{Var}(B)}$	$\frac{\text{Mean}(A)}{\text{Mean}(B)}$
0.2	0.1	0.1	5	0.0098 / 0.0163	0.0038 / 0.0155	0.2416 / 0.2291
0.2	0.1	0.1	20	0.0040 / 0.0090	0.0040 / 0.0089	0.2020 / 0.2096
0.2	0.1	0.1	30	0.0027 / 0.0062	0.0027 / 0.0061	0.2002 / 0.2048
0.2	0.1	0.5	5	0.0240 / 0.0224	0.0211 / 0.0159	0.2538 / 0.2805
0.2	0.1	0.5	20	0.0070 / 0.0376	0.0070 / 0.0343	0.2021 / 0.2571
0.2	0.1	0.8	5	0.0320 / 0.0901	0.0258 / 0.0459	0.2789 / 0.4102
0.2	0.1	0.8	20	0.0079 / 0.0497	0.0079 / 0.0409	0.2022 / 0.2936
0.2	0.5	0.1	5	0.0322 / 0.0515	0.0222 / 0.0299	0.2999 / 0.3471
0.2	0.5	0.1	20	0.0071 / 0.0380	0.0071 / 0.0311	0.2023 / 0.2834
0.2	0.8	0.1	5	0.0344 / 0.0817	0.0229 / 0.0306	0.3075 / 0.4260
0.2	0.8	0.1	20	0.0079 / 0.0723	0.0079 / 0.0496	0.2024 / 0.3506 †
0.01	0.1	0.1	5	0.0090 / 0.0284	0.0018 / 0.0128	0.0949 / 0.1350 †
0.01	0.1	0.1	20	0.0023 / 0.0098	0.0007 / 0.0070	0.0500 / 0.0629

†Note that the bias, introduced by neglecting cases with zero marginals, can be considerable.

where prob(ρ_j) is the probability of obtaining the particular **N** yielding estimated value ρ_j. The kernel used was

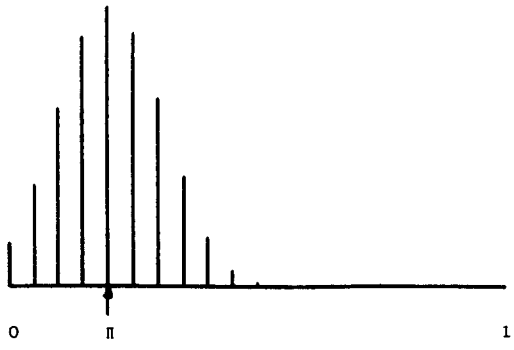
$$K(x) = \begin{cases} \frac{1}{h}\left(1 - \frac{x}{h}\right) & \text{for } |x| < |h|, \\ 0 & \text{else,} \end{cases}$$

with spread parameter $h = 0.2$. (Note that this seems to be an ideal application for the kernel method. Unlike other applications, because it is merely being used to illustrate and compare the general shape of curves, the choice of h is not critical.) Each vertical bar in Figs. 2–5 gives the value of the kernel estimate at that π value. It is apparent from these representative examples that $\hat{\pi}^A$ is the better of the two estimators.

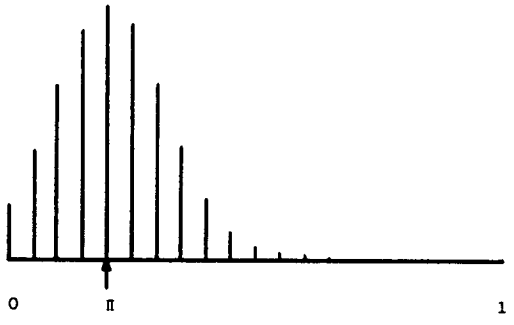
7. CONCLUSION

The simple binomial (or multinomial) estimate of class sizes obtained from the perfect classification of the test set can be improved if we also make use of the very accurate estimates of imperfect classification rates obtained by applying the inexpensive imperfect classifier to a large set of objects. In this paper we have considered two ways in which this extra information may be utilised. One of these estimators ($\hat{\pi}^B$) has some obvious shortcomings: It can lie outside the range [0, 1] and can even take infinite values. The latter point means that the estimator has infinite variance.

Infinite variance might mean that the estimator is of little value—or it might mean that variance is an inappropriate comparison criterion. In the large sample case, for instance, the

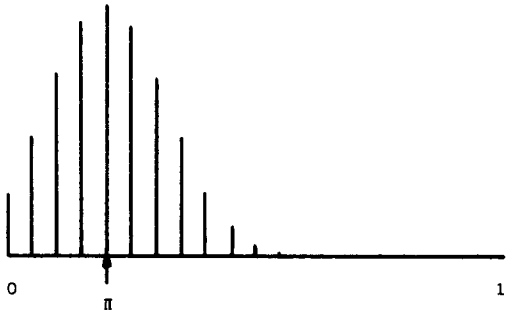


(a)

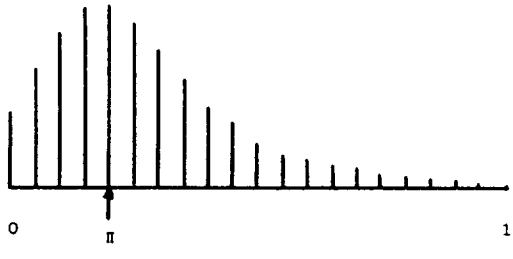


(b)

Fig. 2. A kernel estimate of the distribution of (a) $\hat{\pi}^A$ and (b) $\hat{\pi}^B$ for $\pi = 0.2$, $\phi_1 = 0.1$, $\phi_2 = 0.1$, and $n = 5$.

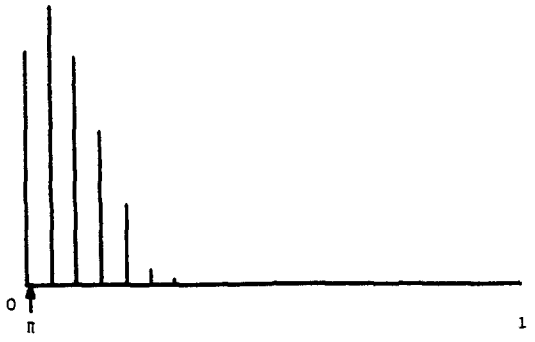


(a)

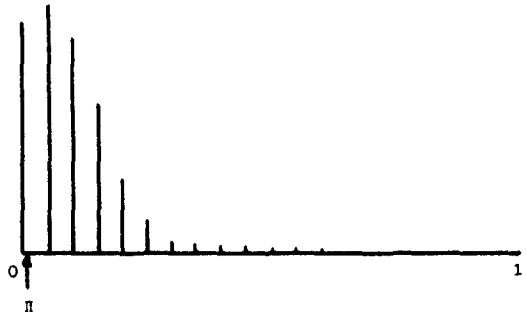


(b)

Fig. 3. A kernel estimate of the distribution of (a) $\hat{\pi}^A$ and (b) $\hat{\pi}^B$ for $\pi = 0.2$, $\phi_1 = 0.1$, $\phi_2 = 0.5$, and $n = 20$.

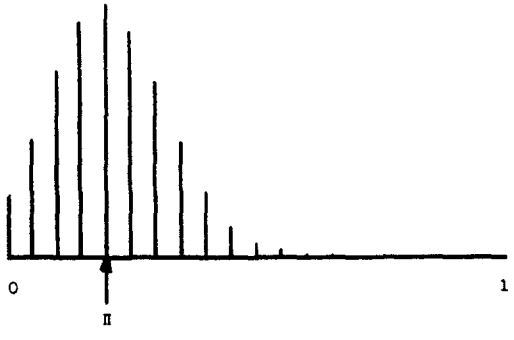


(a)



(b)

Fig. 4. A kernel estimate of the distribution of (a) $\hat{\pi}^A$ and (b) $\hat{\pi}^B$ for $\pi = 0.01$, $\phi_1 = 0.1$, $\phi_2 = 0.1$ and $n = 20$.



(a)

(b)

Fig. 5. A kernel estimate of the distribution of (a) $\hat{\pi}^A$ and (b) $\hat{\pi}^B$ for $\pi = 0.2$, $\phi_1 = 0.5$, $\phi_2 = 0.1$, and $n = 20$.

probability of taking an infinite value might be exceedingly small. We explored this case by studying how small fluctuations of the observed $\hat{\phi}_i$ values away from the true ϕ_i affected the estimators $\hat{\pi}^A$ and $\hat{\pi}^B$. In all cases it seems that the maximum departure of $\hat{\pi}^A$ from the true π is less than the maximum departure of $\hat{\pi}^B$ from the true π. Thus, even for small fluctuations (the vastly more probable case) in the large sample case, $\hat{\pi}^A$ seems superior—as judged by this maximum departure criterion.

Finally, we studied the small sample distributions of $\hat{\pi}^A$ and $\hat{\pi}^B$, eliminating obviously unreasonable $\hat{\pi}^B$ estimates to see if those that remained had better distributional properties than $\hat{\pi}^A$. This seemed not to be the case.

In conclusion, $\hat{\pi}^A$ seems the better estimator, at least as judged by the criteria considered in this paper.

REFERENCES

1. I. Bross, Misclassification in 2 × 2 tables. *Biometrics* **10**, 478–486 (1954).
2. W. G. Cochran, *Sampling Techniques*. John Wiley, New York (1977).
3. W. E. Deming, An essay on screening, or on two-phase sampling, applied to surveys of a community. *Int. Stat. Rev.* **45**, 29–37 (1977).
4. D. J. Hand, *Discrimination and Classification*. John Wiley, Chichester (1981).
5. D. J. Hand, *Kernel Discriminant Analysis*. Research Studies Press, Letchworth (1982).
6. D. J. Hand, in preparation (1985).
7. A. Tenenbein, A double sampling scheme for estimating from binomial data with misclassifications. *J. Am. Stat. Assoc.* **65**, 1350–1361 (1970).
8. A. Tenenbein, A double sampling scheme for estimating from misclassified binomial data: sample size determination. *Biometrics* **27**, 935–944 (1971).
9. A. Tenenbein, A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. *Technometrics* **14**, 187–202 (1972).
10. B. S. White and K. R. Castleman, Estimating cell populations. *Pattern Recognition* **5**, 365–370 (1981).