THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 5, May–June 1981
Printed in U.S.A.

The Effects of Misclassification Error on the Estimation of Several Population Proportions

By J. D. HEALY

(Manuscript received October 21, 1980)

Assume that each item from a set of items is classified into one of several categories. We then use the proportion of items classified into a category to estimate the true proportion of items from the category in the population. This article models the effect of misclassification error on the estimate of the true proportion. We discuss two conditions which can be used to determine the adequacy of a classifier. We present an optimal classification algorithm which can be used when the joint distribution of the variables on which classifications are based is known separately for items from each category.

I. INTRODUCTION

Let us suppose that we observe a set of items which can be split into several distinct categories. Each item is measured and classified by some device into one of these various categories. However, the classified category for an item and the true category may not be the same, i.e., the device may make a misclassification error. The observed proportion of items in a category is then used to estimate the true proportion.

The preceding scenario often occurs in quality control[1] and medical research.[2,3] In quality control, individual manufactured items from a sample or lot are often classified by a mechanical device as defective or not and the proportion of defectives in the sample is then used to estimate the proportion of defectives from the entire process. In medical research, the items are people and the idea is to estimate the proportion of people with various diseases. In a Bell system example, the items would be phone calls and the categories would be busies, completed calls, reorders, etc. An automated device would attempt to determine the true category for each call. The output of the device would then be the estimated proportion of calls in each category. This

last example motivated the analysis contained in this article.[4] Note that the problems discussed here are quite different from the traditional classification problem,[5] where the goal is to maximize the probability of correctly classifying each item. In all the above examples, errors from misclassification can have a serious effect.

In this article, we attack two separate problems. In the first problem, we assume we have little or no control over the internal design of the classifier; all that we have is an estimate of the probability of classifying items from each category into each of the other categories. The object is to develop a simple way to specify how good the classifier must be. Also, we should indicate to the designer the direction in which improvements are necessary. The second problem handles the case when we do have control over the design of the classifier. In this case, we assume that object is to design the classifier so that the effects of misclassification error are minimized. In this article, we are concerned only with a classifier's ability to estimate proportions, i.e., our loss function is entirely different than the usual loss function.

This article is organized in the following way. Section II introduces notation and explains the effects of misclassification error on the estimated proportion of items in a category. Section III discusses the case when we have little or no control over the design of the classifier. In Section IV, we discuss the case when we have control over the classifier. The resulting minimization problem involves a function that is quadratic in the probabilities of classifying items.

II. EFFECTS OF MISCLASSIFICATION

Let m be the number of categories. All vectors in this paper are m-dimensional column vectors, and matrices are m by m matrices. Also let $\mathbf{p}$, $\mathbf{p}^*$, and $\hat{\mathbf{p}}$ be the m-dimensional vectors of the true probabilities of items from the various categories, the probabilities of classifying items into various categories, and the observed proportion of items actually classified into different categories, respectively. Let $A = (a_{ij})$ be the m by m misclassification matrix which contains the conditional probabilities of classifying items into different categories. For example, a_{12} would be the probability of classifying an item from category 2 as an item from category 1. Each column of A sums to one. The diagonal elements of A are the probabilities of correct classification, and the off-diagonal elements give the probabilities of misclassification. A perfect classifier would have $A = I$, the identity matrix. By the law of total probability, $\mathbf{p}^* = A\mathbf{p}$.

In measuring the effectiveness of $\hat{\mathbf{p}}$ as an estimator of $\mathbf{p}$, we use the matrix of mean-squared errors. The matrix of mean-squared errors contains the mean-squared error of the individual terms and also the

cross product terms which indicate how the different errors are related. The m by m matrix of mean-squared errors (MMSE) is

$$E[(\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})' | \mathbf{p}] = \text{cov}(\hat{\mathbf{p}} | \mathbf{p}) + [E(\hat{\mathbf{p}} | \mathbf{p}) - \mathbf{p}][E(\hat{\mathbf{p}} | \mathbf{p}) - \mathbf{p}]', \quad (1)$$

where "cov" means the covariance matrix. The diagonal of the covariance matrix measures precision of an estimator while the diagonal of the second term in (1) is the bias squared.

We assume that n items are to be classified. The expected number of items from each category is $n\mathbf{p}$. If items are classified independently, the distribution of $n\hat{\mathbf{p}}$ will be a multinomial distribution with parameters $A\mathbf{p}$ and sample size n. From Ref. 6, we obtain the $\text{cov}(\hat{\mathbf{p}} | \mathbf{p})$,

$$\text{cov}(\hat{\mathbf{p}} | \mathbf{p}) = [D^* - A\mathbf{p}\mathbf{p}'A']/n, \quad (2)$$

where D^* is a diagonal matrix with diagonal equal to $\mathbf{p}^*$. The $\text{cov}(\hat{\mathbf{p}} | \mathbf{p})$ can be separated into two parts, one part which is the covariance matrix if the classifier were perfect, and the second part which is an adjustment in the covariance matrix because the classifier is not perfect:

$$\text{cov}(\hat{\mathbf{p}} | \mathbf{p}) = [D - \mathbf{p}\mathbf{p}']/n + [D^* - D - A\mathbf{p}\mathbf{p}'A' + \mathbf{p}\mathbf{p}']/n,$$

where D is a diagonal matrix with diagonal equal to $\mathbf{p}$.

Since $E(\hat{\mathbf{p}} | \mathbf{p}) = A\mathbf{p}$, the bias term in (1) becomes

$$[E(\hat{\mathbf{p}} | \mathbf{p}) - \mathbf{p}][E(\hat{\mathbf{p}} | \mathbf{p}) - \mathbf{p}]' = (A - I)\mathbf{p}\mathbf{p}'(A - I)'. \quad (3)$$

Note that this term is not divided by n. Putting the above statements together, (1) becomes

$$E[(\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})' | \mathbf{p}] = (D - \mathbf{p}\mathbf{p}')/n + (D^* - D - A\mathbf{p}\mathbf{p}'A' + \mathbf{p}\mathbf{p}')/n + [(A - I)\mathbf{p}\mathbf{p}'(A - I)']. \quad (4)$$

This equation includes the sampling-error effect (first term on right) and the effect of a misclassification error (other terms on right).

III. SPECIFYING ACCURACY WITH LITTLE CONTROL OVER THE CLASSIFIER

Assume that our information on a classifier is confined to the matrix A, i.e., someone else is responsible for the design of the classifier. We may influence the form of A but we have no control over the actual functioning of the classifier. Suppose we have a good idea of how large the mean-squared errors for each category estimate [the diagonal elements of (4)] can be for the application of the classifier. We need several guidelines on the form of A which will insure that the classifier and its resulting mean-squared errors are adequate. Clearly, we do not

want to put constraints on every element of A. Also, we cannot tell the designer of the classifier that the mean-squared errors of his classifier are inadequate without providing some guidance on how to improve the classifier.

Intuitively, we would like to pick A so that the bias term in (4) disappears. We cannot pick A so that the bias term in (4) disappears for every $\mathbf{p}$. For each A, however, there is some value of $\mathbf{p}$ for which the bias term disappears; in fact, A tends to produce $\hat{\mathbf{p}}$, which are collapsed toward the value of $\mathbf{p}$ for which the bias term disappears. A reasonable strategy is to pick A so that the bias term disappears for our "best guess" for $\mathbf{p}$ which we denote by $\mathbf{p}_0$. This means A should be picked so that $(A - I)\mathbf{p}_0 \approx 0$, i.e., $\mathbf{p}_0$ is approximately an eigenvector of A with eigenvalue 1. There are many A which satisfy $(A - I)\mathbf{p}_0 \approx 0$, but which have large mean-squared errors for values of $\mathbf{p}$ near $\mathbf{p}_0$. To ensure that mean-squared errors are small for values of $\mathbf{p}$ near $\mathbf{p}_0$, we must additionally require that the a_{ii} (diagonal elements of A) be reasonably large; note that if the eigenvector condition is nearly satisfied, the requirement on the a_{ii} may, in many cases, be quite loose.

These two conditions: (1) $(A - I)\mathbf{p}_0 \approx 0$, and (2) a_{ii} large, are generally easy to check. Assume we have a set of s items which we know contain $s\mathbf{p}_0$ items from the respective categories. These items are classified and are the results used to estimate A. The first condition states that the number of items classified into a category is roughly equal to $s\mathbf{p}_0$. If this condition is not originally satisfied, the designer can usually satisfy it by adjusting several thresholds which determine where the classifier places items. The second condition says that the classifier cannot misclassify a high proportion of items from any one category. The designer is then told which categories do not satisfy this condition.

We now show that these two conditions can be justified analytically when we assume that $\mathbf{p}$ has an underlying Dirichlet distribution. That is, we now allow $\mathbf{p}$ to vary, for example, with environment. The Dirichlet distribution is the natural multivariate generalization of the beta distribution and it is the conjugate prior for the multinomial distribution. We pick the parameters of the Dirichlet distribution so that

$$E(\mathbf{p}) = \mathbf{p}_0, \tag{5}$$

$$\text{cov}(\mathbf{p}) = -\mathbf{p}_0\mathbf{p}_0'/(v + 1) + D_0/(v + 1), \tag{6}$$

where D_0 is a diagonal matrix whose ith diagonal element is the ith element of $\mathbf{p}_0$, and v is a parameter that indicates how spread out the Dirichlet distribution is. These parameters, $\mathbf{p}_0$ and v, can be chosen so that the resulting Dirichlet distribution models the expected environ-

mental variability of $\mathbf{p}$. If we take expected values of the terms in (4) using (5) and (6), we obtain

$$E(\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})' = D_0^*/n - AD_0A'/n(v+1) - A\mathbf{p}_0\mathbf{p}_0'A'v/n(v+1) + (A-I)\mathbf{p}_0\mathbf{p}_0'(A-I)'v/(v+1) + (A-I)D_0(A-I)'/(v+1), \quad (7)$$

where D_0^* is a diagonal matrix whose diagonal equals $A\mathbf{p}_0$. This equation incorporates the effects of varying $\mathbf{p}$. If n and v are at all large, the fourth term in (7) will be important. If $(A-I)\mathbf{p}_0 \approx 0$ holds, this term will drop out. The fifth term is minimized if the a_{ii} are large. Since the second and third terms are generally unimportant [they are divided by $n(v+1)$ which should be large] and since the first term is present even with a perfect classifier, satisfying the two conditions will minimize the effects of misclassification. In short, we have presented a way to require the A matrix to be "near" the identity matrix without putting constraints on each and every element of A.

IV. DESIGNING A CLASSIFIER THAT MINIMIZES MEAN-SQUARED ERROR

Assume that our job is to develop a classifier that minimizes the effects of mean-squared error. More specifically, assume we measure a vector of variables ($\mathbf{x}$) on each item. Regions R_i are defined such that if $\mathbf{x} \in R_i$, we classify the item into category i. We want to define the R_i that minimizes a weighted sum of the mean-squared errors for the various categories, i.e., that minimizes

$$\text{tr}[Q\,E(\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})'], \quad (8)$$

where Q is a known positive-diagonal matrix, tr is the trace operator, and $E(\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})'$ is defined by (7). The matrix Q is just a weighting factor that can be used to emphasize the important categories. Minimizing (8) is equivalent to minimizing

$$\text{tr}(QABA' - AC), \quad (9)$$

where

$$B = \frac{n-1}{n(v+1)}(D_0 + \mathbf{p}_0\mathbf{p}_0'v), \quad (10)$$

$$C = \left(\frac{2}{(v+1)}(D_0 + \mathbf{p}_0\mathbf{p}_0'v) - \frac{1}{n}\mathbf{p}_0(1, 1, \cdots, 1)\right)Q. \quad (11)$$

We handle a more general case by allowing B to be any known symmetric, nonnegative definite matrix, and C any known matrix.

Equation (9) is interesting because it is quadratic in A. The usual Bayes multiple decision rule[7] is a special case of (9), since it is the rule

that minimizes (9) when $B = \mathbf{0}$, and C is a diagonal matrix with diagonal equal to the vector of prior probabilities. Classical discriminant analysis[5] is a special case of the Bayes multiple decision rule for which the distribution of $\mathbf{x}$ when the item comes from category i is multivariate normal with mean μ_i and covariance matrix Σ.

Let $f_i(\mathbf{x})$ be the density of $\mathbf{x}$ if $\mathbf{x}$ is measured on an item from category i. The optimal classification algorithm is given in the following theorem:

Theorem 1: Equation (9) will be minimized if $\mathbf{x}$ is classified into the ith category when the ith element of

$$(f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_k(\mathbf{x}))(2BAQ' - C) \tag{12}$$

is the smallest.

(*Proof*: See appendix.)

In general, applying Theorem 1 should be quite difficult since A has to be solved for in (12). We now discuss the two-category case and then specialize to the case when $\mathbf{x}$ has a multivariate normal distribution. We give a simple iterative procedure to calculate the required quantities for this last case.

For the two-category case, Theorem 1 reduces to the following corollary.

Corollary 1: If there are only two categories of measurements, then eq. (9) will be minimized if $\mathbf{x}$ is placed into category 1 if

$$f_1(\mathbf{x})/f_2(\mathbf{x}) > K, \tag{13}$$

where

$$\begin{aligned} K = [&2b_{22}(q_{11} + q_{22})a_{21} - 2b_{21}(q_{11} + q_{22})a_{12} \\ &+ 2(b_{21}q_{11} - b_{22}q_{22}) + (c_{22} - c_{21})] \\ /[&2b_{11}(q_{11} + q_{22})a_{12} - 2b_{12}(q_{11} + q_{22})a_{21} \\ &+ 2(b_{12}q_{22} - b_{11}q_{11}) + (c_{11} - c_{12})], \end{aligned} \tag{14}$$

and b_{ij}, q_{ii}, and c_{ij} are elements of B, Q, and C, respectively.

Let us now assume that, if $\mathbf{x}$ is an observation from category i ($i = 1$ or 2), it has a multivariate normal distribution with known mean vector μ_i and known covariance matrix Σ. Then (13) becomes

$$\log(f_1(\mathbf{x})/f_2(\mathbf{x})) = \mathbf{x}'\Sigma^{-1}(\mu_1 - \mu_2) - 1/2(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) > \log K.$$

As in classical discriminant analysis,[5] the distribution of $\log(f_1(\mathbf{x})/f_2(\mathbf{x}))$ is normal with mean $\alpha/2$ or $-\alpha/2$ when $\mathbf{x}$ comes from categories 1 or 2, respectively, where

$$\alpha = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2). \tag{15}$$

In either case, the variance of $\log(f_1(\mathbf{x})/f_2(\mathbf{x}))$ is α. Therefore, when $\mathbf{x}$ comes from category 1,

$$[\log(f_1(\mathbf{x})/f_2(\mathbf{x})) - \alpha/2]/\sqrt{\alpha}$$

has a standard normal distribution. Similarly,

$$[\log(f_1(\mathbf{x})/f_2(\mathbf{x})) + \alpha/2]/\sqrt{\alpha}$$

has a standard normal distribution when $\mathbf{x}$ comes from category 2. The misclassification probabilities can be defined in terms of a standard normal random variable,

$$a_{21} = P\left(Z < \frac{\log K - \alpha/2}{\sqrt{\alpha}}\right), \tag{16}$$

$$a_{12} = P\left(Z > \frac{\log K + \alpha/2}{\sqrt{\alpha}}\right), \tag{17}$$

where Z has a standard normal distribution. Using (14), (16), and (17) the values of a_{12}, a_{21}, and K may be calculated iteratively:

(1) Let $K = 1$ and calculate a_{12} and a_{21} using (16) and (17);
(2) Obtain a new value of K by substituting a_{12} and a_{21} into K; and
(3) Repeat the entire process.

In summary, assume we are trying to minimize $\mathrm{tr}[QE(\hat{p} - p)(\hat{p} - p)']$, where Q is a known weighting matrix. Also assume we have only two categories and that if $\mathbf{x}$ comes from category i, it has a multivariate normal distribution with mean vector μ_i and covariance matrix Σ. The parameter α may be calculated using (15). The B and C matrices should be calculated using (10) and (11). Equation (13) gives the decision rule for classification into one of the two categories, where the values of K, a_{12}, and a_{21} are obtained in an iterative manner using (14), (16), and (17).

To apply any of the preceding theory, some prior knowledge of the distribution of $\mathbf{p}$ is required to estimate $\mathbf{p}_0$ and v. If the parameters μ_1, μ_2, and Σ are unknown, they may be estimated from a sample of data with the usual sample means and pooled covariance matrix. An estimator of the parameter α could then be calculated using (15) with the estimates of μ_1, μ_2, and Σ substituted into (15). The algorithm discussed above could then be used to generate A and K where the estimator of α is used in (16) and (17) instead of α. The properties of the procedure when estimators of the parameters are used require further study.

V. SUMMARY

This article presents a model that incorporates the effects of mis-

classification error. Several guidelines are presented which can be used to determine if a given classifier is adequate. For the case when the classifier is yet to be designed, we have given an optimal classification algorithm. We discuss the two-category case in detail.

APPENDIX

Proof of Theorem 1

Let R_i^0 be the regions that result if the theorem is applied. Let A_0 be the resulting misclassification matrix. Let R_i^1 and A_1 be the corresponding elements for some other decision rule. Now consider (9) evaluated at A_1 minus (9) evaluated at A_0:

$$\begin{aligned}\operatorname{tr}(QA_1BA_1' - QA_0BA_0' - A_1C + A_0C) \\ = \operatorname{tr}(A_1 - A_0)B(A_1 - A_0)'Q + 2\operatorname{tr}(A_1 - A_0)BA_0'Q \\ - \operatorname{tr}(A_1 - A_0)C. \quad (18)\end{aligned}$$

The first term on the right of (18) is nonnegative since B and Q are nonnegative definite. We still have to show that the rest of (18) is nonnegative. Consider

$$2\operatorname{tr}(A_1 - A_0)BA_0'Q - \operatorname{tr}(A_1 - A_0)C = \operatorname{tr}A_1E - \operatorname{tr}A_0E, \quad (19)$$

where $E = 2BA_0'Q - C$. Let e_{ij} be the i,jth element of E. Equation (19) now becomes

$$\begin{aligned}\operatorname{tr}(A_1E) - \operatorname{tr}(A_0E) &= \sum_{i,j} \int \phi_1(i|\mathbf{x}) f_j(\mathbf{x})\, d\mathbf{x} \cdot e_{ji} \\ &\quad - \sum_{k,m} \int \phi_0(k|\mathbf{x}) f_m(\mathbf{x})\, d\mathbf{x} \cdot e_{mk} \\ &= \sum_{i,k} \int \phi_1(i|\mathbf{x})\phi_0(k|\mathbf{x}) \\ &\quad \cdot \left[\sum_j f_j(\mathbf{x}) e_{ji} - \sum_m f_m(\mathbf{x}) e_{mk}\right] d\mathbf{x}, \quad (20)\end{aligned}$$

where

$$\phi_0(i|\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in \mathbf{R}_i^0, \\ 0, & \mathbf{x} \notin R_i^0, \end{cases}$$

$$\phi_1(i|\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in R_i^1, \\ 0, & \mathbf{x} \notin R_i^1. \end{cases}$$

Since $\phi_0(k|\mathbf{x})$ will be zero whenever $[\sum_j f_j(\mathbf{x}e_{ji} - \sum_m f_m e_{mk}]$ is negative, (20) is always nonnegative. Q.E.D.

REFERENCES

1. W. M. Wooding, "A Source of Bias in Attributes Testing, and a Remedy," J. Quality Technol., *11*, No. 4 (October 1979), pp. 169–76.
2. P. Armitage, *Statistical Methods in Medical Research*, New York: Wiley, 1971.
3. T. Colton, *Statistics in Medicine*, Boston: Little, Brown, and Co., 1974, pp. 87–92.
4. J. D. Healy, unpublished work.
5. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958, Chap. 6.
6. N. L. Johnson and S. Kotz, *Discrete Distributions*, New York: Wiley, 1969, pp. 281–91.
7. T. S. Ferguson, *Mathematical Statistics A Decision Theoretic Approach*, New York: Academic, 1973, Chap. 6.