Optimal Sample Allocation for Normal Discrimination and Logistic Regression under Stratified Sampling

by

Tzu-Cheg Kao
Uniformed Services University
of the Health Sciences

and

George P. McCabe
Purdue University

Technical Report #86–28

Department of Statistics
Purdue University

July 1986
Revised June 1990

△

Optimal Sample Allocation for Normal Discrimination and Logistic Regression under Stratified Sampling

by

Tzu-Cheg Kao
Uniformed Services University
of the Health Sciences

and

George P. McCabe*
Purdue University

ABSTRACT

For two multivariate normal populations with a common covariance matrix and stratified sampling, we consider two methods of estimation—Fisher's linear discriminant function and logistic regression. Intuition suggests that taking half of the observations from each population is a reasonable design choice. Based upon minimizing the expected error regret, asymptotic optimal sample allocations are found. The results indicate that the differences in the expected error regret for optimal versus balanced allocation are generally quite small. It is recommended that equal sample sizes for the two populations be used for these problems.

KEY WORDS: Optimal sampling plan; Stratified sampling, Fisher's linear discriminant function; Logistic regression.

---

*Tzu-Cheg Kao is Assistant Professor, Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814. George P. McCabe is Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907.

△

1 INTRODUCTION

Assume that an individual belongs to one of two populations H_0 and H_1, indicated by $Y = 0$ and $Y = 1$, respectively. Let $P\{Y = 0\} = \pi_0$ and $P\{Y = 1\} = \pi_1$, where $\pi_0 + \pi_1 = 1$, and $0 < \pi_1 < 1$. Let X be a k-dimensional random vector which is measured on each individual. We assume that the distribution of X given $Y = i$ is multivariate normal with mean μ_i and covariance matrix Σ, and denote the corresponding density by $f_i(x)$, where $i = 0$ or 1.

Suppose that an individual with an observed value for X, say x, is to be classified into one of the two populations. The rule which minimizes the expected probability of misclassification assigns x to population H_1, if $\pi_1 f_1(x) \geq \pi_0 f_0(x)$ and to population H_0, otherwise.

If the consequences of misclassification are the same for each population then the expected probability of misclassification is equivalent to the expected loss. In the more general case, l_i is the loss associated with misclassification of an individual from population i, and the loss is zero for a correct classification. Here, we let $\pi_0' = 1-\pi_1' = l_0\pi_0/(l_0\pi_0+l_1\pi_1)$ and note that minimizing expected loss for the general case is equivalent to minimizing the expected misclassification probability with (π_0, π_1) replaced by (π_0', π_1').

Under a zero-one loss function, this rule is the Bayes rule. See Chapter 6 of Anderson (1984) for details. Let $P\{Y = i|x\}$ be the posterior probability that $Y = i$ given x, where $i = 0$ or 1. Then $\pi_1 f_1(x) \geq \pi_0 f_0(x)$ if and only if $P\{Y = 1|x\} \geq P\{Y = 0|x\}$. From the normality assumption, it follows that

$$1 - P\{Y = 0|x\} = P\{Y = 1|x\} = \frac{\exp(\alpha + \delta' x)}{1 + \exp(\alpha + \delta' x)}, \tag{1.1}$$

where

$$\alpha = \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}\left(\mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0\right), \tag{1.2}$$

and

$$\delta = \Sigma^{-1}(\mu_1 - \mu_0).$$

Let $\beta = (\alpha, \delta')'$. We wish to estimate the unknown parameter β using stratified sampling, i.e., $\{X_1, \ldots, X_{n_1}\}$, a random sample of size n_1 is taken from population H_1 and $\{X_{n_1+1}, \ldots, X_{n_1+n_0}\}$, a random sample of size n_0 is taken from population H_0. Let $n = n_0 + n_1$. The sample allocation problem is viewed as follows. For a fixed value of n, let π^* denote the proportion of observations taken from H_1, i.e. $\pi^* = n_1/n$. The design problem is to choose a value of π^*.

Anderson (1972) suggests that $\pi^* = .5$ is a reasonable choice for logistic regression. He writes, "It is conjectured that for a given total sample size n, samples with balance give better estimates, on average, than those with imbalance."

We study this problem for two different estimators of β. The first is the maximum likelihood estimator (MLE). Here we find the MLE's of μ_1, μ_0 and Σ by maximizing the likelihood function

$$\ell(\mu_1, \mu_0, \Sigma) = \prod_{i=1}^{n_1} f_1(x_i) \cdot \prod_{i=n_1+1}^{n_1+n_0} f_0(x_i),$$

where $f_i(x)$ is the multivariate normal density with parameters μ_i and Σ. The MLE of β is then found using the relationships given in (1.2). This approach will be referred to as *normal discrimination*. Here we assume that π_1 is known. If π_1 is unknown, the MLE of β still can be found from (1.2) given an independent estimator of π_1.

The *logistic regression* estimator is found by maximizing

$$\ell(\beta) = \prod_{i=1}^{n_1} \frac{\exp(\alpha + \delta' x_i)}{1 + \exp(\alpha + \delta' x_i)} \cdot \prod_{i=n_1+1}^{n} \frac{1}{1 + \exp(\alpha + \delta' x_i)}.$$

Note that this method of estimation does not require the assumption that X given Y is multivariate normal. It is sufficient to assume that the posterior given by (1.1) has the

logistic form. Details regarding estimation for this model can be found in Efron (1975), Blyth and McLachlan (1978) and McLachlan (1980).

The expected error regret (EER) is the difference between the expected misclassification probability using a specified estimation procedure and the misclassification probability that would be obtained if all parameters were known. Asymptotic expressions for EER are used to determine the optimal sample allocation for any given parametric configuration and to compare the optimal allocation with $\pi^* = .5$.

In a typical application one of the populations represents individuals who have a disease such as a particular type of cancer. The other population represents similar individuals who do not have the disease. If the disease is relatively rare then a random sample from the combined population will contain very few individuals with the disease. Under such circumstances it is common practice to use the stratified sampling procedure described above. Our results indicate that equal sample sizes are a very good choice in a wide variety of circumstances.

2 PRELIMINARY THEOREMS

In many practical applications, π_1 is not known but can be estimated from an independent random sample for which the X's are not measured. We assume that $Y_1, \ldots Y_m$ are i.i.d. Bernoulli random variables with parameter π_1 and that this sample is independent of the stratified sample. Let $\hat{\pi}_1 = \overline{Y}$. We further assume that $\lim_{n,m\to\infty}(n/m) = R$, where R is finite.

Let $\Delta = \{(\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 - \mu_0)\}^{\frac{1}{2}}$. Since our results are invariant with respect to linear transformations of X, it is sufficient to consider a standardized form of the problem, as in Efron (1975). In this form, $\mu_1 = (\Delta/2)e_1, \mu_0 = -(\Delta/2)e_1$ and $\Sigma = I$. Here, $e_1 = (1, 0, \ldots, 0)'$ and I is the $k \times k$ identity matrix. Note that in this form $\alpha = \log(\pi_1/(1-\pi_1))$

and $\delta = \Delta e_1$, i.e., $\beta = (\log(\pi_1/\pi_0), \Delta e_1')'$.

The following theorem gives the asymptotic distribution of the MLE produced from the normal discrimination procedure for π_1 unknown or known.

Theorem 2.1. Suppose π_1 is unknown and it is estimated by $\hat{\pi}_1$ from an independent sample as described above. The normal discrimination procedure gives the MLE estimator $\hat{\beta}_M^n$, that is consistent and satisfies

$$\sqrt{n}\left(\hat{\beta}_M^n - \beta\right) \xrightarrow{L} N_{k+1}(0, \Sigma_M),$$

where

$$\Sigma_M = \frac{1}{\pi^*(1-\pi^*)} \begin{bmatrix} \frac{\Delta^2}{4} + \frac{R\pi^*(1-\pi^*)}{\pi_1\pi_0} & -\frac{\Delta}{2}(1-2\pi^*) & 0 \ldots 0 \\ -\frac{\Delta}{2}(1-2\pi^*) & 1+2\Delta^2\pi^*(1-\pi^*) & 0 \ldots 0 \\ 0 & 0 & W \ldots 0 \\ \vdots & \vdots & \ddots \\ 0 & 0 & 0 \ldots W \end{bmatrix},$$

and $W = 1 + \Delta^2\pi^*(1-\pi^*)$. If π_1 is known, the above results hold with $R = 0$.

The proof of the above theorem uses ideas similar to those used in Efron (1975) and follows directly from Theorem 3.3.1 and its corollary in Kao (1982).

The logistic regression procedure gives an estimator $\tilde{\beta}_L^n$, that converges almost surely to $\beta^* = (\alpha^*, \delta')'$, where $\alpha^* = \log(\pi^*/(1-\pi^*))$ and $\delta = \Delta e_1$. For technical details, see Theorem 3.2.4 by Kao (1982).

Recall that $\alpha = \log(\pi_1/\pi_0)$. Thus, the constant term in the logistic regression estimator is asymptotically biased. This bias is easily removed by adding $\log[(\pi_1/(1-\pi_1))/(\pi^*/(1-\pi^*))]$ to the first component of $\tilde{\beta}_L^n$. We denote the adjusted estimator by $\hat{\beta}_L^n$. Note that if π_1 is unknown, its value $\hat{\pi}_1$ from an independent sample is used in place of π_1 in the adjustment.

The following theorem gives the asymptotic properties for the adjusted estimator, $\hat{\beta}_L^n$.

Theorem 2.2. Suppose π_1 is unknown and it is estimated by $\hat{\pi}_1$ from an independent sample as described above. The adjusted estimator $\hat{\beta}_L^n$ is consistent and satisfies

$$\sqrt{n}(\hat{\beta}_L^n - \beta) \xrightarrow{L} N_{k+1}(0, \Sigma_L),$$

where

$$\Sigma_L = \frac{1}{\pi^*(1-\pi^*)} \times \begin{bmatrix} \frac{A_2}{A_0A_2-A_1^2} - 1 + \frac{R\pi^*(1-\pi^*)}{\pi_1\pi_0} & \frac{-A_1}{A_0A_2-A_1^2} & 0 & \dots & 0 \\ \frac{-A_1}{A_0A_2-A_1^2} & \frac{A_0}{A_0A_2-A_1^2} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{A_0} & & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & \dots & \frac{1}{A_0} \end{bmatrix},$$

and

$$A_i = \frac{\exp(-\Delta^2/8)}{\sqrt{2\pi}} \int \frac{x^i \exp(-x^2/2)}{\pi^* \exp(\Delta x/2) + (1-\pi^*)\exp(-\Delta x/2)} dx, \qquad i = 0, 1, 2.$$

If π_1 is known, the above results hold with $R = 0$.

The proof of the above theorem uses ideas similar to those used in Efron (1975) and follows directly from Corollaries 3.2.4.1 and 3.2.4.2 in Kao (1982).

3 MINIMIZATION OF THE EXPECTED ERROR REGRET

First, we define the error rate of a classification rule as the probability that it misclassifies a random individual. The true error rate is equal to

$$\pi_0 \int_{\{\alpha+\delta' x>0\}} f_0(x)dx + \pi_1 \int_{\{\alpha+\delta' x\leq 0\}} f_1(x)dx$$

Suppose $(\alpha, \delta')' = \beta$ is estimated by an estimator, say $\hat{\beta}^n$, satisfying the following asymptotic condition

$$\sqrt{n}(\hat{\beta}^n - \beta) \xrightarrow{L} N_{k+1}(0, V),$$

where $V = (v_{ij})$ $i, j = 0, 1, \ldots, k$. Under these conditions, Efron (1975) gives the following asymptotic expression (ignoring terms of order less than $1/n$) for the expected error rate using the estimator $\hat{\beta}^n$,

$$\pi_1 \Phi(-d_1) + \pi_0 \Phi(-d_0) + \frac{\pi_1 \varphi(d_1)}{2\Delta n} \left[v_{00} - \frac{2\lambda}{\Delta} v_{01} + \frac{\lambda^2}{\Delta^2} v_{11} + v_{22} + \ldots + v_{kk} \right] \tag{3.1}$$

where $d_1 = \Delta/2 + \lambda/\Delta, d_o = \Delta/2 - \lambda/\Delta, \lambda = \log(\pi_1/\pi_0), \Delta = \left[(\mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0)\right]^{\frac{1}{2}}$, and $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and cumulative distribution function, respectively.

The first two terms of (3.1) give the error rate for the case where all parameters are known. Therefore, we define the (asymptotic) expected error regret as

$$\frac{\pi_1 \varphi(d_1)}{2\Delta n} \left[v_{00} - \frac{2\lambda}{\Delta} v_{01} + \frac{\lambda^2}{\Delta^2} v_{11} + v_{22} + \ldots + v_{kk} \right] \tag{3.2}$$

The optimal sample allocations for normal discrimination and logistic regression are discussed in detail in subsections 3.1 and 3.2 respectively. The main results are given in Theorems 3.1, 3.2 and 3.3.

3.1 Normal Discrimination

Assume that π_1 is unknown. From Theorem 2.1 and (3.2), the asymptotic expected error regret for normal discrimination is

$$\frac{\pi_1 \varphi(d_1)}{2\Delta n} \left\{ h(\pi^*) + 2\lambda^2 + (k-1)\Delta^2 + \frac{R}{\pi_0 \pi_1} \right\}, \tag{3.3}$$

where

$$h(\pi^*) = \frac{\left(\frac{\Delta}{2} + \frac{\lambda}{\Delta}\right)^2}{\pi^*} + \frac{\left(\frac{\Delta}{2} - \frac{\lambda}{\Delta}\right)^2}{1 - \pi^*} + \frac{(k-1)}{\pi^*(1-\pi^*)}. \tag{3.4}$$

If π_1 is known, the asymptotic expected error regret is obtained by taking $R = 0$ in the above expression. Since the function $h(\pi^*)$ contains all of the dependence of the expected

error regret on π^*, it is sufficient to minimize this quantity to determine optimal values of π^*.

From (3.4), it can be seen that particular difficulties arise when $k = 1$ and $\lambda = \pm\Delta^2/2$. Under these conditions, $h(\pi^*)$ is minimized at either $\pi^* = 0$ or $\pi^* = 1$. Recall that the validity of (3.1) depends upon the asymptotic normality of $\hat{\beta}$. Since this condition would not be satisfied with $\pi^* = 0$ or 1, we exclude these special cases from further consideration.

Using the expression for the expected error regret given by (3.3) we can find values of π^* that minimize this quantity. Note that the results are based on asymptotic considerations that require both n_1 and n_2 to approach infinity at the same rate. The following two theorems give the basic results.

Theorem 3.1 The value $\pi^* = .5$ minimizes the asymptotic expected error regret for normal discrimination under each of the following conditions: (a) $k \to \infty$, (b) $\Delta \to \infty$, (c) $\Delta \to 0$, (d) $\pi_1 = .5$.

Proof. As k gets large, $h(\pi^*)$ is approximately $(k-1)/[\pi^*(1-\pi^*)]$. As Δ gets large $h(\pi^*)$ is approximately $\Delta^2/[4\pi^*(1-\pi^*)]$. As $\Delta \to 0, h(\pi^*)$ is approximately $\lambda^2/[\Delta^2\pi^*(1-\pi^*)]$. For $\pi_1 = .5$, $\lambda = 0$ and $h(\pi^*) = a/[\pi^*(1-\pi^*)]$. These expressions are all minimized by $\pi^* = .5$.

Other cases are covered by the following theorem.

Theorem 3.2 For normal discrimination when $\pi_1 \neq .5$ and $\lambda \neq \pm\Delta^2/2$ with $k = 1$, the optimal sample allocation is $\pi^* = (a - \sqrt{a(a-2\lambda)})/(2\lambda)$ where $a = \Delta^2/4 + \lambda^2/\Delta^2 + \lambda + k - 1$.

Proof. The result follows by setting the derivative of $h(\pi^*)$ equal to zero and taking the root corresponding to $0 < \pi^* < 1$.

3.2 Logistic Regression

Assume that π_1 is unknown. From Theorem 2.2 and (3.2), it follows that the asymp-

totic expected error regret is

$$\frac{\pi_1\varphi(d_1)}{2\Delta n}\left\{g(\pi^*)+R/(\pi_0\pi_1)\right\}, \tag{3.5}$$

where

$$g(\pi^*)=\frac{1}{\pi^*(1-\pi^*)}\left\{\frac{A_2+2(\lambda/\Delta)A_1+\lambda^2A_0/\Delta^2}{A_0A_2-A_1^2}-1+\frac{k-1}{A_0}\right\}. \tag{3.6}$$

and the A_i are defined in Theorem 2.2. If π_1 is known, the asymptotic expected error regret can be obtained by setting $R=0$ in the above expression.

The following theorem gives the optimal value of π^* for some special cases.

Theorem 3.3 The value $\pi^*=.5$ minimizes the asymptotic expected error regret for logistic regression under each of the following conditions: (a) $k\to\infty$, (b) $\Delta\to 0$, (c) $\pi_1=.5$.

Proof. It is sufficient to restrict attention to the function $g(\pi^*)$. (a) As k gets large, $g(\pi^*)$ is approximately $(k-1)/(\pi^*(1-\pi^*)A_0)$. The result follows from examination of the derivatives. (b) As Δ approaches 0, the terms A_0, A_1 and A_2 approach 1, 0, and 1, respectively. In this case $g(\pi^*)$ is approximately $\lambda^2/[\Delta^2\pi^*(1-\pi^*)]$ which is minimized at $\pi^*=.5$. (c) The condition $\pi_1=.5$ is equivalent to $\lambda=0$. Under this condition, we can write $g(\pi^*)$ as

$$g(\pi^*)=\frac{1}{\pi^*(1-\pi^*)}\left[\frac{A_2}{A_0A_2-A_1^2}-1\right]+\frac{k-1}{\pi^*(1-\pi^*)A_0}.$$

We noted in part (a) that the second term is minimized when $\pi^*=.5$. Since $1/[\pi^*(1-\pi^*)]$ is minimized when $\pi^*=.5$, it remains to show that $A_2/(A_0A_2-A_1^2)-1$ is also minimized when $\pi^*=.5$. Since $A_1^2\geq 0$, this expression is bounded below by $A_0^{-1}-1$ where A_0 is evaluated at $\pi^*=.5$. Note that $A_1^2=0$ if and only if $\pi^*=.5$. Thus, the bound is achievable and the expression is minimized when $\pi^*=.5$. Combining the above and noting that $A_0<1$ when $\pi^*=.5$ gives the desired result.

Note that the expression for $g(\pi^*)$ given by (3.6) involves π^* in the A_i terms. Since these terms involve π^* in a complex way, it is not possible to find the value of π^* that minimizes $g(\pi^*)$ analytically. On the other hand, the expression can be evaluated numerically. Results of these types of calculation lead to the conjecture that $\pi^* = .5$ as Δ approaches infinity.

4 Numerical Computations

Optimal values of π^* for selected values of $k, \pi_1 = .50(.05).95$ and $\Delta = 2.0$ are given in Table 1. Qualitatively similar results hold for different values of Δ. For normal discrimination, these values are obtained from Theorems 3.1 and 3.2. For logistic regression, the function $g(\pi^*)$ given by (3.6) was evaluated for $\pi^* = .500(.005).995$ and the minimizing value of π^* in this set is reported. Computational details are given in Kao (1982).

Following the definition of $h(\pi^*)$ in section 3.1, it is noted that technical difficulties arise with the asymptotics when $k = 1$ and $\lambda = \pm\Delta^2/2$. For $\Delta = 2$ this corresponds to $\pi_1 = .88$ (or $\pi_0 = .12$). The effect can be seen in Table 1 where π^* is large for values of π_1 near this point.

Values for $\pi_1 < .5$ can be obtained by interchanging the populations H_0 and H_1. Note that $\pi^* > .5$ whenever $\pi_1 > .5$, suggesting that the sample size should be larger in the population having the higher prior probability. This conclusion easily follows from Theorem 3.2 for normal discrimination and from Theorem 4.2.2 in Kao (1982) for logistic regression.

For both normal discrimination and logistic regression, the asymptotic expected error regret goes to zero as n^{-1}. Thus, the error rate using these estimators approaches the true error rate as long as n_1 and n_2 approach infinity. By choosing the optimal π^* the coefficient of n^{-1} in the expression (3.2) for the asymptotic expected error regret is minimized. In

what follows, we refer to this coefficient as the rate constant.

By comparing the rate constants for different choices of π^*, we can investigate the sensitivity of our results with respect to these choices. Specifically, we compare the rate constant for the choice $\pi^* = .5$ to that for the optimal π^*.

Table 2 gives the differences between the rate constants for $\pi^* = .5$ and the optimal π^* for $\Delta = 2$ and normal discrimination and logistic regression. Note that all differences are zero for $\pi_1 = .5$ since $\pi^* = .5$ is optimal for this case. As might be expected from Theorems 3.1 and 3.3, the differences become smaller as k increasses. Similar results hold for different values of Δ.

5 CONCLUSIONS

For given values of π_1, Δ and k we have shown how to compute the optimal sample allocation for both normal discrimination and logistic regression. Table 1 gives optimal values of π^* for selected cases. Although in some cases, particularly when k is small, the optimal π^* is not particularly close to .5. Table 2 indicates that very little is lost in terms of expected error rate by choosing $\pi^* = .5$. Note that the entries in these tables are divided by n to obtain the differences in the asymptotic expected error rates.

In many practical situations the value of Δ is not known in advance of the data collection although reasonable guesses may be available. The same may be true of π_1 and, in some cases, even k. Since very little is lost by choosing $\pi^* = .5$, we recommend this choice for most practical applications.

It is important to note that the above conclusions are based upon consideration of the expected error rate. When using these procedures for classification, the intercept α in addition to the coefficient vector δ must be estimated. For some problems, however, interest may be confined to the estimation of δ. In such cases, it is easy to see that the

arguments given in this paper can be modified to show that $\pi^* = .5$ is optimal in the sense that it minimizes the standard errors of the parameter estimator of any linear combination of the components of δ.

Situations where the cost of an observation depends upon which population is being sampled are easily accomodated within the present framework. Suppose that c_i represents the cost of an observation from population i. If the total cost $T = c_0 n_0 + c_1 n_1$ is fixed, then minimization of the asymptotic expected error regret is achieved by choosing $n_1 = \pi^* T/(c_0 + (c_1 - c_0)\pi^*)$ and $n_0 = (1-\pi^*)T/(c_0 + (c_1 - c_0)\pi^*)$. When $\pi^* = .5$, this reduces to $n_1 = n_0 = T/(c_0 + c_1)$. In practice, the n_i obtained by these formulas are not necessarily integers. The optimal sample allocation is obtained by checking $[n_i]$ and $[n_i] + 1$.

REFERENCES

Anderson, J.A. (1972), "Separate Sample Logistic Discrimination," *Biometrika*, **59**, 19–35.

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, (2nd ed.), New York: Wiley.

Blyth, K. and McLachlan, G.J. (1978), "The Biases Associated with Maximum Likelihood Methods of Estimation of the Multivariate Logistic Risk Function," *Communications in Statistics*, A7, 877–890.

Efron, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of American Statistical Association*, **70**, 892–898.

Kao, T.C. (1982), Maximum Likelihood Discrimination and Logistic Regression, Ph.D. Thesis, Purdue University.

McLachlan, G.J. (1980), "A Note on Bias Correction in Maximum Likelihood Estimation with Logistic Discrimination," *Technometrics*, **22**, 621–627.