Choice of Mixture and Stratified Sampling for Normal and Logistic Discrimination

by

Tzu-Cheg Kao
University of Wisconsin

George P. McCabe
Purdue University

Technical Report #86-33

Department of Statistics
Purdue University

July 1986

Choice of Mixture and Stratified Sampling for Normal and Logistic Discrimination

by

Tzu-Cheg Kao and George P. McCabe*

ABSTRACT

Relative efficiencies of logistic regression to normal discrimination in the stratified sampling case, and mixture sampling to stratified sampling for logistic regression and normal discrimination are derived. These results can be used to decide whether mixture sampling or stratified sampling is preferred.

1. Introduction

Suppose that a k-dimensional vector X can only arise from one of two normal populations H_0 and H_1 with mean μ_h and common covariance Σ, where $h = 0$ with probability π_0 if X is from H_0, and $h = 1$ with probability π_1 otherwise, where $\pi_1 + \pi_0 = 1$ and $0 \leq \pi_0 < 1$. Denote the probability density function of x given h by $P_h(\cdot)$. If an individual with an observed value for X, say x, is to be classified into one of the two populations. Define $L9x, \beta) = \beta_0 + \delta' x$, where

$$\begin{aligned} \beta_0 &= \log(\frac{\pi_1}{\pi_0}) - \frac{1}{2}(\mu_1'\sigma^{-1}\mu_1 - \mu_0'\sigma^{-1}\mu_0), \\ \delta &= \Sigma^{-1}(\mu_1 - \mu_0), \\ \Sigma &= (\sigma^{ij})_{k\times k'} \\ \mu_1 &= (\mu_{1i})_{k\times 1}, \\ \mu_0 &= (\mu_{0i})_{k\times 1}. \end{aligned} \tag{1.1}$$

The classification rule, i.e., Fisher's "linear discriminant function", is to assign x to population H_1 if $L(x, \beta) > 0$ and to population H_0 otherwise. It is well-known that the rule is optimal in the sense that it minimizes the expected probability of misclassification (Anderson, 1958). Under zero-one loss function, it is a Bayes rule w.r.t. π_1 too. Let $P\{H_i|x\}$ be the posterior probability of H_i given X, where $i = 0$ or 1. Note that $L(x, \beta) \geq 0$ if and only if

$$\pi_1 P_1(x) \geq \pi_0 P_0(x). \tag{1.2}$$

Also, we observe that (1.2) holds if and only if

$$P\{H_1|x\} \geq P\{H_0|x\}. \tag{1.3}$$

---

* Tzu-Cheg Kao is Assistant Professor, Department of Mathematics, University of Wisconsin-Oshkosh, WI 54901. George P. McCabe is Professor, Head of Statistical Consulting, Department of Statistics, Purdue University, West Lafayette, IN 47907. This work was partially supported by the University of Wisconsin-Oshkosh Faculty Development Program CAS Grant.

Key Words: Classification, Efficiency, Sampling plan, and Discriminant Analysis.

we note that

$$P\{H_1|x\} = \exp(\beta_0 + \delta' x)/[1 + \exp(\beta_0 + \delta' x)],$$
$$P\{H_0|x\} = 1 - P\{H_1|x\}, \tag{1.4}$$

where $(\beta_0, \delta') = \beta$; is a $(k+1)$-dimensional parameter vector. It is easy to show that (1.4) holds if and only if

$$\pi_1 P_1(x)/[\pi_0 P_0(x)] = \exp(\beta_0 + \delta' x) \tag{1.5}$$

holds.

Now, we consider estimation of the unknown parameter β. First of all, we focus on two types of sampling: (1) mixture sampling, that is, $\{X_1, h_1), \ldots, (X_n, h_n)\}$, a random sample of size n is taken from a mixture of the two populations; (2) stratified sampling, that is, $\{X_1, \ldots, X_{n_1}\}$, 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 $n = n_1 + n_0$. For each type of sampling, we consider two methods for finding estimators of β. Throughout this article, we define $\pi^* = n_1/n$.

The first approach is to use the maximum likelihood estimation (MLE). This will be referred to as "normal discrimination". We can find the MLE of μ_1, μ_0, and Σ. Therefore, by (1.1), the MLE of β can be found if π_1 is known. If π_1 is unknown, then we need some conditional like (A) (stated in section 2) in order to obtain the MLS of β.

the second approach is to use the so-called logistic regression. Here, we denote it as "logistic discrimination". That is, to find an estimator of β (i.e., M-estimator of β) by maximizing $\ell_M(\beta)$ (or $\ell_X(\beta)$) with respect to β if we use mixture (or stratified) sampling, where

$$\ell_M(\beta) = \prod_{i=1}^{n} (P_1 x_i)^{h_i} (P_0 x_i)^{1-h_i},$$
$$\ell_S(\beta) = \prod_{i=1}^{n} P_1 x_i \cdot \prod_{i=n_1+1}^{n_1+n_0} P_0 x_i, \text{ where}$$
$$P_{1X} = \exp(\beta_0 + \delta' x)/[1 - \exp(\beta_0 + \delta' x)], \text{ and}$$
$$P_{0X} = 1 - P_{1x}.$$

1.2 Motivation

Efron ([5] in 1975) compared logistic discrimination approach with normal discrimination approach based only on mixture sampling. He just made his comment on stratified sampling in the following quotation:

"Most frequently, the sample size n_1 (for the first group) and n_0 (for the second group) are set by the statistician and are not random variables at all." (Efron [5] in 1975, p. 898).

Please note that he mentioned it and did not really compare mixture sampling with stratified sampling. This comparison has not yet actually been studied rigorously. The lack of any rigorous study motivates me to investigate this issue.

Suppose that there is a choice between mixture sampling and stratified sampling. Generally speaking, the goal of this article is to work out answers for the following two problems:

(1) Using the normal discrimination approach, under what other conditions is the mixture sampling preferred to the stratified sampling and vice versa?

(2) Using the logistic discrimination approach, under what other conditions is the mixture sampling preferred to the stratified sampling and vice versa?

1.3 Summary

In section 2, we give 4 Lemmas. They provide background to obtain the main result. In sections 3 and 4, we study the relative efficiencies of mixture to stratified sampling for logistic and normal discrimination respectively. Some relative efficiency plots are provided. We then make choice between mixture and stratified sampling for each approach. Based on our findings, we give two suggestions in section 5.

2. Preliminary Lemmas

This section defines the expected error rate (EER), and the relative efficiency measure. Four lemmas are given on the EER for normal or logistic discrimination under mixture or stratified sampling. Firstly, we consider the case where π_1 is unknown or no prior information available. In order to estimate the π_1, we impose on the following extra condition:

(A) *Let another sample of size m be taken from the mixture of two populations, independent of the stratified sample of size n, and* $\frac{n}{m} \to R$, *a non-negative finite constant, when* $n \to \infty$.

We define $\hat{\pi}_1$ = proportion of individuals which came from population H_1 in the random sample of size m. It is easy to see that $\hat{\pi}_1$ is the MLE of π_1 satisfying

$$\hat{\pi}_1 \xrightarrow{\text{a.s.}} \pi_1 \tag{2.1}$$

$$\text{and } \sqrt{m}\{\hat{\pi}_1 - \pi_1\} \xrightarrow{\text{L}} N(0, \pi_1\pi_0) \tag{2.2}$$

Secondly, we restate the definition of "expected error rate" (Efron 1975). Suppose that under some method of estimation (say, M), $\hat{\beta}^{(n)}$ satisfies

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

where $V = (v_{ij})$. From Theorem 1 in Efron (1975), the expected error rate is approximately

$$EER(M) = \frac{\pi_1\phi(d)}{2\Delta n} \cdot \left[v_{00} - \frac{2\lambda}{\Delta}v_{01} + \frac{\lambda^2}{\Delta^2}v_{11} + v_{22} + \cdots + v_{pp}\right], \tag{2.3}$$

ignoring terms of order less than $1/n$, where $d = \Delta/2 + \lambda/\Delta$,

$$\phi(d) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{d} e^{-t^2/2} dt, \quad \lambda = \log(\pi_1/\pi_0), \text{ and } \Delta = \{(\mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0)\}.$$

Also, we see that the expected error rate is unchanged under linear transformations of the data. Without loss of generality, we consider the standardized cases where

$$\mu_1 = (\Delta/2)e_1, \quad \mu_0 = -(\Delta/2)e_1, \quad \text{and } \Sigma = I. \tag{2.4}$$

Hereafter, we will write EER instead of EER(M) if no ambiguity occurs.

Thirdly, we restate the "relative efficiency" measure (Efron, 1975) which is used through this article. Now let A, B be two methods of estimation of β. We define the relative efficiency of B to A, denoted $EFF_k(A, B)$, by

$$EFF_k(A, B) = \frac{EER(A)}{EER(B)}, \tag{2.5}$$

where the expected error rates, $EER(A)$ and $EER(B)$, are defined as in (2.3). We note that $EFF_k(A, B) < 1$ if and only if A is more efficient than B. Also, we define the asymptotic relative efficiency of B to A, denoted $EFF_\infty(A, B)$, by $EFF_\infty(A, B) = \lim_{k \to \infty} EFF_k(A, B)$. If no ambiguity occurs, we may write $EFF_k(A, B), EFF_\infty(A, B)$ simply as EFF_k, EFF_∞ respectively.

Recall that we can restrict attention to the standardized situation (2.4), without loss of generality. Let

$$\begin{aligned}
Q_1^* &= \frac{\Delta^2}{4} + \lambda(1 - 2\pi^*) + \frac{\lambda^2}{\Delta^2}[1 + 2\Delta^2 \pi^*(1 - \pi^*)] + \frac{R\pi^*(1 - \pi^*)}{\pi_0 \pi_1}, \\
Q_2^* &= 1 + \Delta^2 \pi^*(1 - \pi^*), \\
Q_3^* &= \frac{A_2^* + 2\lambda A_1^*/\Delta + \lambda^2 A_0^*/\Delta^2}{A_0^* A_2^* - A_1^{*2}} - 1 + \frac{R\pi^*(1 - \pi^*)}{\pi_0 \pi_1}, \\
\text{and } Q_4^* &= 1/A_0^*, \\
W^* &= \Delta^2/4 + \lambda(1 - 2\pi^*) + \lambda^2/\Delta^2[1 + 2\Delta^2 \pi^*(1 - \pi^*)].
\end{aligned}$$

where Q_1, Q_2, Q_3, Q_4 and W are obtained from $Q_1^*, Q_2^*, Q_3^*, Q_4^*$ and W^* by replacing π_1 by π^* in the above formulas.

Now we state the following four lemmas. Proofs can be found in the Appendix.

Lemma 1. Assume that normal discrimination with stratified sampling is used. (i) If π_1 is unknown, then

$$EER = \pi_1 \phi(d)[h(\pi^*) + R/(\pi_0 \pi_1)]/[2\Delta n],$$

where

$$h(\pi^*) = \{\Delta^2/4 + \lambda(1-2\pi^*) + \lambda^2[1+2\Delta^2\pi^*(1-\pi^*)]/\Delta^2 + (k-1)[1+\Delta^2\pi^*(1-\pi^*)]\}/[\pi^*(1-\pi^*)].$$

(ii) If π_1 is known, then

$$EER = \pi_1\phi(d)h(\pi^*)/[2\Delta n].$$

Lemma 2. Assume that logistic regression with stratified sampling is used. (i) If π_1 is unknown, then

$$EER = \pi_1\phi(d)[h(\pi_1, \pi^*) + R/(\pi_0\pi_1)]/[2\Delta n],$$

where

$$h(\pi_1, \pi^*) = \{[A_2^* + 2\lambda A_1^*/\Delta + \lambda^2 A_0^*/\Delta^2]/(A_0^*A_2^* - A_1^{*2}) - 1 + (k-1)/A_0^*\}/[\pi^*(1-\pi^*)].$$

(ii) If π_1 is known, then

$$EER = \pi_1\phi(d)h(\pi_1, \pi^*)/[2\Delta n].$$

Lemma 3. Assume that normal discrimination with mixture sampling is used.

(i) If π_1 is unknown, then

$$EER = \frac{\pi_1\phi(d)}{2\Delta n}\{h(\pi_1) + 1/(\pi_0\pi_1)\},$$

where

$$h(\pi_1) = \frac{1}{\pi_1\pi_0}\{\Delta^2/4 + \lambda(1-2\pi_1) + \lambda^2/\Delta^2[1+2\Delta^2\pi_1\pi_0] + (k-1)(1+\Delta^2\pi_1\pi_0)\}.$$

(ii) if π_1 is known, then

$$EER = \pi_1\phi(d)h(\pi_1)/[2\Delta n].$$

Lemma 4. If one uses logistic regression and mixture sampling, then

$$EER = \frac{\phi(d)}{2\Delta\pi_0 n} \cdot \left\{ \frac{A_2 + 2A_1\lambda/\Delta + \lambda^2 A_0/\Delta^2}{A_0A_2 - A_1^2} + \frac{k-1}{A_0} \right\}.$$

Note: The above is true whether π_1 is known or unknown.

3. Mixture vs. Stratification for Logistic Discrimination

In this section, we assume that we want to use the logistic discrimination approach. Depending on π_1 being known or unknown, Theorem 1 gives the relative efficiency of mixture sampling to stratified sampling. Two corollaries are derived. We will then use them to compare mixture sampling versus stratified sampling.

Theorem 1. (i) If π_1 is unknown, then the relative efficiency of mixture sampling to stratified sampling for logistic regression is

$$EFF_k = \frac{Q \cdot EFF_1 + (k-1)EFF_\infty}{Q + (k-1)}, \tag{3.1}$$

where

$$EFF_1 = \frac{\pi_0 \pi_1 Q_3^*}{\pi^*(1-\pi^*)(Q_3 - R + 1)},$$

$$EFF_\infty = \frac{\pi_1 \pi_0 Q_4^*}{\pi^*(1-\pi^*)Q_4},$$

and

$$Q = \frac{Q_3 - R + 1}{Q_4}.$$

(ii) If π_1 is known, then (3.1) is valid with $R = 0$.

Proof: We consider two cases:

Case (i): π_1 is unknown. From Lemmas 4 and 2, we see that

$$EFF_k = \frac{\pi_1 \pi_0 (Q_3^* + (k-1)Q_4^*)}{\pi^*(1-\pi^*)(Q_3 + 1 - R + (k-1)Q_4)}.$$

We can rewrite it as a weighted average of the relative efficiencies when $k = 1$ and $k = \infty$. Hence (i) is proved.

Case (ii): π_1 is known. From Lemmas 2 and 4, we see that EFF_k is obtained by taking $R = 0$ in Case (i). □

Anderson (1972) suggested that $\pi^* = 1/2$ is intuitively reasonable. The conjuncture has been somewhat confirmed by Kao and McCabe (1986). Here, we are interested in making comparison between both sampling schemes when $\pi^* = 1/2$. It is easy to see the following.

Corollary 1.1. Assume that $\pi^* = 1/2$. (i) If π_1 is unknown, then

$$EFF_k = \frac{Q \cdot EFF_1 + (k-1)EFF_\infty}{Q + (k-1)}, \tag{3.2}$$

where

$$EFF_1 = \frac{4\pi_1\pi_0\left\{\frac{A_2^*+(\lambda^2/\Delta^2)A_0^*}{A_0^*A_2^*} - 1 + \frac{R}{4\pi_1\pi_0}\right\}(A_0A_2 - A_1^2)}{(A_2 + 2\lambda A_1/\Delta + \lambda^2 A_0/\Delta^2)},$$

$$EFF_\infty = 4\pi_1\pi_0 A_0/A_0^*,$$

and

$$Q = \frac{(A_2 + 2\lambda A_1/\Delta + \lambda^2 A_0/\Delta^2)A_0}{(A_0A_2 - A_1^2)}.$$

(ii) If π_1 is known, then (3.2) is valid with $R = 0$.

This corollary implies that $EFF \leq 1$, if we take $\pi^* = 1/2$. In other words, in this case, the stratified sampling is preferred to mixture sampling. Figures 1 and 2 give plots of relative efficiencies with respect to π_1, for $\pi^* = 1/2$, $\pi_1 = .5(.005).995$ and $\Delta = 2, 4$.

In case $\pi^* \neq 1/2$, we make comparisons by varying all possible π^* according to π_1. We obtain the following Corollary 1.2. Assume that $\pi^* = \pi_1$. (i) If π_1 is unknown, then

$$EFF_k = \frac{Q \cdot EFF_1 + (k-1)EFF_\infty}{Q + k - 1}, \tag{3.3}$$

where

$$EFF_1 = 1 + \frac{R-1}{Q_3 - R + 1},$$

$$EFF_\infty = 1,$$

and

$$Q = \frac{Q_3 - R + 1}{Q_4}.$$

(ii) If π_1 is known, then (3.3) is valid with $R = 0$.

From Corollary 1.2, we have

$$EFF_1 \left\{\begin{matrix}\leq\\=\\>\end{matrix}\right\} 1 \qquad \text{if} \qquad R \left\{\begin{matrix}\leq\\=\\>\end{matrix}\right\} 1$$

In other words, we have

$$EFF_k \left\{\begin{matrix}\leq\\=\\>\end{matrix}\right\} 1 \qquad \text{if} \qquad R \left\{\begin{matrix}\leq\\=\\>\end{matrix}\right\} 1.$$

Figures 3 and 4 give plots of relative efficiencies with respect to π_1 for $\pi^* = \pi_1, \pi_1 = .5(.005).995$ and $\Delta = 2.4$. If $R \leq 1$, then there exists π^* such that EFF_k (in the formula (3.1)) is less than or equal to 1. This can be justified by Theorem 3.2 (Kao and McCabe 1986). Here we note that $EFF_k < 1$ if $R < 1$, and $EFF_k \leq 1$ with strict inequality for some π_1 if $R = 1$. In this case, stratified sampling is preferred. Now if $R > 1$, then the mixture sampling plan provides more information on π_1 than the additional sample of size m. Therefore, mixture sampling is preferred.

4. Mixture vs. Stratification for Normal Discrimination

In this section, we assume that we want to use the normal discrimination approach. Theorems 2 and 3 give the relative efficiencies of mixture sampling to stratified sampling for the case when π_1 is unknown and known respectively. Based on the two derived corollaries and Theorem 3, we suggest a choice between the two sampling schemes.

4.1 π_1 unknown.

Theorem 2. If π_1 is unknown, then the relative efficiency of mixture sampling to stratified sampling for normal discrimination is

$$EFF_k = \frac{Q \cdot EFF_1 + (ki-1)EFF_\infty}{Q+k-1}, \tag{4.1}$$

where

$$EFF_1 = \frac{\pi_0\pi_1}{\pi^*(1-\pi^*)} \frac{Q_1^*}{Q_1+1-R},$$
$$EFF_\infty = \frac{\pi_0\pi_1}{\pi^*(1-\pi^*)} \cdot \frac{Q_2^*}{Q_2}$$

and

$$Q = \frac{Q_1+1-R}{Q_2}.$$

Proof: From Lemmas 1 and 3, we see that

$$EFF_k = \frac{\pi_0\pi_1(Q_1^* + (k-1)Q_2^*)}{\pi^*(1-\pi^*)\{Q_1+1-R+(k-1)Q_2\}}.$$

We can rewrite this expression as a weighted average of the relative efficiencies when $K=1$ and $K=\infty$. Hence the theorem is shown □

Here, the same comment just after Theorem 1 follows. It is straightforward to obtain

Corollary 2.1. For $\pi^* = 1/2$, the expression (4.1) holds with

$$EFF_1 = \frac{\pi_1\pi_0\left[\Delta^2 + 4(1+\Delta^2/2)\lambda^2/\Delta^2 + R/(\pi_1\pi_0)\right]}{\Delta/4 + \lambda(1-2\pi_1) + (1+2\Delta^2\pi_1\pi_0)\lambda^2/\Delta^2 + 1},$$
$$EFF_\infty = \frac{\pi_0\pi_1(4+\Delta^2)}{1+\Delta^2\pi_1\pi_0},$$
$$Q = \frac{\Delta^2/4 + \lambda(1-2\pi_1) + \lambda^2(1+2\Delta^2\pi_1\pi_0)/\Delta^2 + 1}{1+\Delta^2\pi_1\pi_0}.$$

From corollary 2.1, it is easy to show that if $\pi^* = 1/2$ then $EFF_\infty \leq 1$, where the inequality can be strict for $\pi_1 \neq 1/2$. Therefore, in this case, stratified sampling is

preferred to mixture sampling for normal discrimination. Figures 5 and 6 give plots of relative efficiencies with respect to π_1, for $\pi^* = 1/2$, $\pi_1 = .5(.005).995$ and $\Delta = 2, 4$.

For the case $\pi^* \neq 1/2$, again we make comparisons by varying all possible π^* according to π_1.

Corollary 2.2. For $\pi^* = \pi_1$, we have (4.1) with

$$EFF_1 = \frac{Q_1}{Q_1 + 1 - R},$$
$$EFF_\infty = 1,$$

and

$$Q = \frac{Q_1 + 1 - R}{Q_2}.$$

Recall that $R = \lim_{m,n\to\infty} n/m$, in condition (A) of section 2. from the above corollary, it follows that

$$EFF_1 \left\{\begin{matrix}\leqq \\ = \\ \geqq\end{matrix}\right\} 1 \qquad \text{if } R \left\{\begin{matrix}\leqq \\ = \\ \geqq\end{matrix}\right\} 1.$$

that is,

$$EFF_k \left\{\begin{matrix}\leqq \\ = \\ \geqq\end{matrix}\right\} 1 \qquad \text{if } R \left\{\begin{matrix}\leqq \\ = \\ \geqq\end{matrix}\right\} 1.$$

Figures 7 and 8 give plots of relative efficiencies with respect to π_1, for $\pi^* = \pi_1$, $\pi_1 = .5(.005).995$ and $\Delta = 2, 4$. If $R \leq 1$, then there exists π^* such that EFF_k (in the formula (4.1)) is less than or equal to 1. This can be shown from Theorem 3.1 (Kao and McCabe 1968). Here we note that $EFF_k < 1$ if $R < 1$, and $EFF_k \leq 1$ with strict inequality for some π_1 if $R = 1$. In this case stratified sampling is preferred. Now if $R > 1$, then the mixture sampling plan provides more information on π_1 than the additional sample of size m. Therefore, mixture sampling is preferred.

4.2 π_1 known.

Theorem 3. If π_1 is known, then the relative efficiency of mixture sampling to stratified sampling for normal discrimination is

$$EFF_k = \frac{Q \cdot EFF_1 + (k-1)EFF_\infty}{Q + k - 1}, \tag{4.1}$$

where

$$EFF_1 = \frac{\pi_1\pi_0}{\pi^*(1-\pi^*)} \cdot \frac{W^*}{W},$$
$$EFF_\infty = \frac{\pi_1\pi_0}{\pi^*(1-\pi^*)} \cdot \frac{Q_2^*}{Q_2},$$

and

$$Q = \frac{W}{Q_2}.$$

Proof: From Lemmas 1 and 3, we have

$$EFF_k = \frac{\pi_1\pi_0[W^* + (k-1)Q_2^*]}{\pi^*(1-\pi^*)[W + (k-1)Q_2]}.$$

We can rewrite this as a weighted average of the relative efficiencies when $k = 1$ and $k = \infty$. Hence the theorem is proven. □

From Theorem 3.1 by Kao and McCabe (1986), there exists optimal π^* in (0,1) satisfying

$$\frac{1}{\pi^*(1-\pi^*)}[W^* + (k-1)Q_2^*] \leq \frac{1}{\pi_0\pi_1}(W + (k-1)Q_2),$$

where the inequality can be strict for some π_1. In this case, therefore, we have $EFF_k \leq 1$. In other words, stratified sampling is preferred to mixture sampling.

Figures 9, and 10 give relative efficiencies with respect to π_1, where π^* is optimally obtained as in Theorem 3.1 of Kao and McCabe (1986) for $\pi_1 = .5(.005).995$, and $\Delta = 2, 4$ for $\pi_1 = .5(.005).955$,

5. Suggestions

Suppose that the underlying distributions are multivariate normal with different means and common covariance. Recall that $R = \lim_{n,m\to\infty} n/m$, in condition (A) of section 2. If one has a choice between mixture sampling and stratified sampling, then we suggest the following two decision rules on making a choice between them.

(i) If π_1 is known, then it is better to use stratified sampling.

(ii) If π_1 is unknown, then stratified sampling is preferred to mixture sampling if $R \leq 1$, and mixture sampling is preferred to stratified sampling if $R > 1$.

The above holds no matter which approach (i.e. normal or logistic discrimination) you want to use. It also show us the robustness by using the logistic regression approach.

Appendix

Proof of Lemma 1. (i) Assume that π_1 is unknown. From Theorem 3.3.1 (Kao, 1982), we note that the MLE of $\beta, \hat{\beta}_2$ has asymptotic covariance matrix,

$$\Sigma^*_{L_{(k+1)x(k+1)}} = \frac{1}{\pi^*(1-\pi^*)}$$

$$\times \begin{bmatrix} \frac{\Delta^2}{4} + \frac{R\pi^*(1-\pi^*)}{\pi_1\pi_0} & -\frac{\Delta}{2}(1-2\pi^*) & 0 & \cdots & 0 \\ -\frac{\Delta}{2}(1-2\pi^*) & 1+2\Delta^2\pi^*(1-\pi^*) & 0 & \cdots & 0 \\ 0 & 0 & 1+\Delta^2\pi^*(1-\pi^*) & & \\ \vdots & \vdots & & & \\ 0 & 0 & \cdots & & 1+\Delta^2\pi^*(1-\pi^*) \end{bmatrix} \tag{A.1}$$

The expected error rate is $\frac{\pi_1\phi(d)}{2\Delta n}\ \{h(\pi^*) + \frac{R}{\pi_0\pi_1}\}$, where

$$h(\pi^*) = \frac{1}{\pi^*(1-\pi^*)}\Big\{\frac{\Delta^2}{4} + \lambda(1-2\pi^*) + \frac{\lambda^2}{\Delta^2}[1+2\Delta^2\pi^*(1-\pi^*)] + (k-1)[1+\Delta^2\pi^*(1-\pi^*)]\Big\}.$$

(ii) if π_1 is known, then it follows from Corollary 3.3.1 (Kao, 1982) that the asymptotic covariance of $\hat{\beta}_S$ is obtained by taking $R = 0$ in Σ^*_L. Therefore, the expected error rate is

$$\pi_1\phi(d)h(\pi^*)/[2\Delta n]. \qquad \square$$

Proof of Lemma 2. From Theorem 6.4 (see Kao and McCabe, 1983) the M-estimator of $\beta, \hat{\beta}_{SL}$ exists and has the asymptotic covariance matrix

$$J^*_L = J^{-1} - E_{11}/\{\pi^*(1-\pi^*)\},$$

where

$$J = \pi^*(1-\pi^*)\begin{bmatrix} A^*_0 & A^*_1 & 0 & \cdots & 0 \\ A^*_1 & A^*_2 & 0 & \cdots & 0 \\ 0 & 0 & A^*_0 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & A^*_0 \end{bmatrix}_{(k+1)x(k+1)},$$

and

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

$i = 0, 1, 2$. Therefore

$$J_L^* = \frac{1}{\pi^*(1-\pi^*)} \cdot$$

$$\begin{bmatrix} \frac{A_2^*}{A_0^* A_2^* - A_1^{*2}} - 1 & \frac{-A_1^*}{A_0^* A_2^* - A_1^{*2}} & 0 & \dots & 0 \\ \frac{-A_1^*}{A_0^* A_2^* - A_1^{*2}} & \frac{A_0^*}{A_0^* A_2^* - A_1^{*2}} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{A_0^*} & 0 & \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{A_0^*} \end{bmatrix} \qquad (A.2)$$

(i) Suppose that π_1 is unknown. From theorem 6.6 (see Kao and McCabe, 1983), we note that the adjusted M-estimator of $\beta, \hat{\beta}_{SL}$ has the asymptotic covariance matrix

$$J_L^* + \frac{R}{\pi_0 \pi_1} \cdot E_{11}.$$

The expected error rate is

$$\{h(\pi_1, \pi^*) + R/(\pi_0\pi_1)\}\pi_1\phi(d)/[2\Delta n],$$

where

$$h(\pi_1, \pi^*) = \frac{1}{\pi^*(1-\pi^*)}\{\frac{A_2^* + 2(\lambda/\Delta)A_1^* + \lambda^2 A_0^*/\Delta^2}{A_0^* A_2^* - A_1^{*2}} - 1 + \frac{k-1}{A_0^*}\}.$$

(ii) Suppose that π_1 is known. From theorem 6.5 (see Kao and McCabe, 1983), the adjusted M-estimator of $\beta, \hat{\beta}_{SL}$ has the asymptotic covariance J_L^*. Therefore, the expected error rate is

$$\frac{\pi_1\phi(D)}{2\Delta}(h(\pi_1, \pi^*)). \qquad \square$$

Proof of Lemma 3. (i) Assume that π_1 is unknown. From theorem 3.3.3 (see Kao 1982), we note that the MLE of $\beta, \hat{\beta}_M$ has asymptotic covariance Σ_L which is obtained from the matrix Σ_L^* in (A.1) by replacing π_1 for π^* and taking $R = 1$. Therefore, referring to the proof of Lemma 1, the expected error rate is

$$\frac{\pi_1\phi(d)}{2\Delta n}\{h(\pi_1) + 1/\pi_0\pi_1\},$$

where

$$h(\pi_1) = \frac{1}{\pi_1\pi_0}\{\Delta^2/4 + \lambda(1-2\pi_1) + \lambda^2/\Delta^2[1+2\Delta^2\pi_1\pi_0] + (k-1)(1+\Delta^2\pi_1\pi_0)\}.$$

(ii) If π_1 is known, then it follows from Corollary 3.3.3 (see Kao 1982), that the asymptotic covariance of $\hat{\beta}_M$ is obtained from the matrix Σ_L^* in (A.1) by replacing π_1 for π^* and taking $R = 0$. Hence, referring to the proof of Lemma 1 again, the expected error rate is

$$\pi_1\phi(d)h(\pi_1)/[2\Delta n]. \qquad \square$$

Proof of Lemma 4. from Theorem 5.3 (Kao and McCabe 1983) or the Lemma 3 by Efron (1975), the M-estimator of $\beta, \hat{\beta}_{ML}$ has asymptotic covariance,

$$J_L^{-1} = \frac{1}{\pi_1\pi_0}$$

$$\begin{bmatrix} \frac{A_2}{A_0A_2-A_1^2} & \frac{-A_1}{A_0A_2-A_1^2} & 0 & \cdots & 0 \\ \frac{-A_1}{A_0A_2-A_1^2} & \frac{A_0}{A_0A_2-A_1^2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{A_0} & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{A_0} \end{bmatrix}$$

The result follows from (2.3). $\square$

References

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Efron, B. (1975). The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. *J. Amer. Statist. Assoc.* 70, 892-898.

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

Kao, T. C. and McCabe, G. P. (1983), "Asymptotic Properties of M-estimators with Applications in Logistic Regression", Revised October 1984, Technical Report #83-8, Purdue University, Department of Statistics.

Kao, T. C. and McCabe, G. P. (1986), "Optimal Sample Allocation for Normal Discrimination and Logistic Regression under Stratified Sampling", Technical Report #86-28, Purdue University, Department of Statistics.

Figure Titles and Legends

Figure 1. *Relative efficiencies* for $\pi^* = .5$, $\pi_1 = .5(.005).995$, when $\Delta = 2.0$ *in section* 3. Read below from "↓": EFF_1 for (1) $R = 1.2$, (2) $R = 1.0$, (3) $R = .8$, (4) $R = .6$, (5) $R = .5$, (6) $R = .4$, (7) $R = .2$, (8) $R = 0$; and (9) EFF_∞.

Figure 2. *Relative efficiencies* for $\pi^* = .5$, $\pi_1 = .5(.005).995$, when $\Delta = 4.0$ *in section* 3. Read below from "↓": EFF_1 for (1) $R = 1.2$, (2) $R = 1.0$, (3) $R = .8$, (4) $R = .6$, (5) $R = .5$, (6) $R = .4$, (7) $R = .2$, (8) $R = 0$; and (9) EFF_∞.

Figure 3. *Relative efficiencies* for $\pi^* = \pi_1, \pi_1 = .5(.005).995$, when $\Delta = 2.0$ *in section* 3. Read below from "↓": EFF_1 for (1) $R = 1.2$, (2) $R = 1.0$, (3) $R = .8$, (4) $R = .6$, (5) $R = .5$, (6) $R = .4$, (7) $R = .2$, (8) $R = 0$; and (9) EFF_∞.

Figure 4. *Relative efficiencies* for $\pi^* = \pi_1, \pi_1 = .5(.005).995$, when $\Delta = 4.0$ *in section* 3. Read below from "↓": EFF_1 for (1) $R = 1.2$, (2) $R = 1.0$, (3) $R = .8$, (4) $R = .6$, (5) $R = .5$, (6) $R = .4$, (7) $R = .2$, (8) $R = 0$; and (9) EFF_∞.

Figure 5. *Relative efficiencies* for $\pi^* = 5, \pi_1 = .5(.005).995$, when $\Delta = 2.0$ *in section* 4.1. Read below from "↓": EFF_1 for (1) $R = 1.2$, (2) $R = 1.0$, (3) $R = .8$, (4) $R = .6$, (5) $R = .5$, (6) $R = .4$, (7) $R = .2$, (8) $R = 0$; and (9) EFF_∞.

Figure 6. *Relative efficiencies* for $\pi^* = .5, \pi_1 = .5(.005).995$, when $\Delta = 4.0$ *in section* 4.1. Read below from "↓": EFF_1 for (1) $R = 1.2$, (2) $R = 1.0$, (3) $R = .8$, (4) $R = .6$, (5) $R = .5$, (6) $R = .4$, (7) $R = .2$, (8) $R = 0$; and (9) EFF_∞.

Figure 7. *Relative efficiencies* for $\pi^* = \pi_1, \pi_1 = .5(.005).995$, when $\Delta = 2.0$ *in section* 4.1. Read below from "↓": EFF_1 for (1) $R = 1.2$, (2) $R = 1.0$, (3) $R = .8$, (4) $R = .6$, (5) $R = .5$, (6) $R = .4$, (7) $R = .2$, (8) $R = 0$; and (9) EFF_∞.

Figure 8. *Relative efficiencies* for $\pi^* = \pi_1, \pi_1 = .5(.005).995$, when $\Delta = 4.0$ *in section* 4.1. Read below from "↓": EFF_1 for (1) $R = 1.2$, (2) $R = 1.0$, (3) $R = .8$, (4) $R = .6$, (5) $R = .5$, (6) $R = .4$, (7) $R = .2$, (8) $R = 0$; and (9) EFF_∞.

Figure 9. *Relative efficiencies* for given π_1, the optimal π^* vs. $\pi_1 = .5(.005).995$, when $\Delta = 4.0$ *in section* 4.2. Read below from "↓": EFF_k for (1) $k = 1$, (2) $k = 3$, (3) $k = 5$, (4) $k = 7$, (5) $k = 9$, (6) $k = 11$, (7) $k = 13$, (8) $k = 15$; and (9) $k = 17$, (10) $k = 19$, (11) $k = 21$; and (12) k_∞.

Figure 10. *Relative efficiencies* for given π_1, the optimal π^* vs. $\pi_1 = .5(.005).995$, when $\Delta = 4.0$ *in section* 4.2. Read below from "↓": EFF_k for (1) $k = 1$, (2) $k = 3$, (3) $k = 5$, (4) $k = 7$, (5) $k = 9$, (6) $k = 11$, (7) $k = 13$, (8) $k = 15$; and (9) $k = 17$, (10) $k = 19$, (11) $k = 21$; and (12) k_∞.