September 1968

SOME CONDITIONAL ESTIMATION PROBLEMS WITH APPLICATIONS TO ESTIMATING THE PROBABILITY OF MISCLASSIFICATION*

by

Marilyn Sorum and Robert J. Buehler

Technical Report No. 112

University of Minnesota
Minneapolis, Minnesota

*This research was supported by National Science Foundation Grant GP-6859.

1. INTRODUCTION

Let x denote observations arising from a distribution $f(x; \theta)$ where θ is unknown, and let $R(x)$ denote some known function of the data x. We will be concerned with the estimation from x of functions of the form $\phi(R(x), \theta)$. The problem arises in the theory of classification where ϕ denotes the probability that a classification rule R based on data x will misclassify a future observation from a specified population characterized by an unknown parameter θ. A natural way to construct an estimator of ϕ would be to put $\hat{\phi}(x) = \phi(R(x), \hat{\theta}(x))$ where $\hat{\theta}(x)$ is, say, the maximum likelihood estimator of θ. We shall however approach the problem after the fashion of the standard Cramér-Rao and Rao-Blackwell theory, considering the conditional variance of estimators which are conditionally unbiased, given $R(x)$. Two examples are given, one involving a classification problem for normal distributions.

2. THE CRAMÉR-RAO BOUND

The usual Cramér-Rao theory extends in a straightforward way to include the present case. To simplify the presentation, we will treat the case of n observations from a continuous distribution, putting $\underset{\sim}{x} = (x_1, \ldots, x_n)$. We further suppose that $R(\underset{\sim}{x})$ is such that we can find $(n-1)$ coordinates $y_1, \ldots, y_{n-1}$ giving a "smooth" transformation (one-to-one with Jacobian existing and nowhere equal to 0 or ∞) from $\underset{\sim}{x}$ to $\underset{\sim}{y}$, where $\underset{\sim}{y} = (y_1, \ldots, y_n)$, $y_n = R(\underset{\sim}{x})$.

Theorem 1. Let $\phi(R, \theta)$ be a function such that $\partial\phi/\partial\theta$ exists. Let $g_\theta(y_1, \ldots, y_{n-1}|y_n)$ denote the conditional density of $y_1, \ldots, y_{n-1}$ given y_n, and assume this density satisfies the usual regularity conditions (see, for example, Lehmann [6]). Let $\hat{\phi}(\underset{\sim}{x})$ be any conditionally unbiased estimator (i.e., $E(\hat{\phi}(\underset{\sim}{x})|Y_n = y_n) = \phi(y_n, \theta)$). Then $\hat{\phi}$ satisfies

$$(1) \quad \mathrm{Var}(\hat{\phi}(\underset{\sim}{x})|Y_n = y_n) \geq \{E(S^2|Y_n = y_n)\}^{-1}$$

where

$$(2) \quad S = \partial \log g_\theta/\partial\phi = (\partial \log g_\theta/\partial\theta)(\partial\phi/\partial\theta)^{-1}.$$

We supress the proof, which follows the usual lines. Superficially it appears that the bound (1) depends on the arbitrary choice of coordinates $y_1, \ldots, y_{n-1}$. But this is not actually so, for let $z_1, \ldots, z_{n-1}$ be any other choice and let g^*_θ and S^* be the analogues of g_θ and S. Then $g^*_\theta = Jg_\theta$, where $J = |\partial(y_1, \ldots, y_{n-1})/\partial(z_1, \ldots, z_{n-1})|$, and $S^* = S$, so that S^* and S have the same second moment.

It is possible to avoid the regularity conditions on g_θ and on $\phi(R, \theta)$ by the method of Chapman and Robbins [3]. We will state the result without proof.

Theorem 2. In the notation of Theorem 1, define

$$(3) \quad A = A(\phi, \theta, h, r) = h^{-1}[\phi(r, \theta+h) - \phi(r, \theta)]$$

$$(4) \quad B = B(\phi, \theta, h, r) = A^{-2}h^{-2}\{[\frac{g_{\theta+h}(y_1, \ldots, y_{n-1}|r)}{g_\theta(y_1, \ldots, y_{n-1}|r)}]^2 - 1\}.$$

Then for any conditionally unbiased estimator $\hat{\phi}(\underset{\sim}{x})$ of $\phi(R, \theta)$,

$$(5) \quad \mathrm{Var}(\hat{\phi}|r) \geq \{\inf E[B(\phi, \theta, h, r)|r]\}^{-1}$$

where the infimum is taken over all $h \neq 0$ such that both θ and $\theta+h$ lie in the range of possible θ values.

3. RAO-BLACKWELL THEORY

In order to give the desired modification of the Rao-Blackwell theory, we will require definitions of conditional sufficiency and conditional complete sufficiency.

Definition 1. A statistic $T(x)$ is called conditionally sufficient for θ given $R(x)$ if $(R(x), T(x))$ is sufficient for θ.

To discuss completeness, we will consider separately the discrete case and the absolutely continuous case.

Definition 2 (discrete case). A statistic $T(x)$ is called a conditionally complete sufficient statistic for θ given $R(x)$ if for each value r of R

$$(6) \quad \sum_t f(t)P_\theta(T=t|R=r) = 0 \quad \text{for all} \quad \theta$$

implies

$$(7) \quad f(t) = 0 \quad \text{for all} \quad t \quad \text{such that} \quad P_\theta(T=t|R=r) > 0.$$

Definition 2 (absolutely continuous case). Let $h_\theta(\cdot)$ be the density of a statistic $R(x)$ and let $g_\theta(\cdot|r)$ be the conditional density of a statistic $T(x)$ given $R(x) = r$, defined for all r such that $h_\theta(r) > 0$. $T(x)$ is called conditionally complete sufficient given R if for each r such that $h_\theta(r) > 0$,

$$(8) \quad \int f(t)g_\theta(t|r)dt = 0 \quad \text{for all} \quad \theta$$

implies

$$(9) \quad f(t) = 0 \quad \text{a.e.} \quad P_\theta^{T|R=r}.$$

It is more or less evident from Definition 1 that every sufficient statistic is conditionally sufficient (see Bahadur [2], Theorem 6.4, for a rigorous treatment). It is also true that completeness implies conditional completeness. We give a proof for the discrete case only.

Theorem 3. Let P_θ be a family of probability measures defined on a discrete space $\{x\}$ assigning nonzero probability to each point for each value of θ. If $T(x)$ is a complete sufficient statistic for θ, then $T(x)$ is a conditionally complete sufficient statistic for θ given any statistic $R(x)$.

Proof. Let $\mathcal{T}$, $\mathcal{R}$ and $\mathcal{S}$ denote the range of $T(x)$, $R(x)$ and $(T(x), R(x))$ respectively. Every point of $\mathcal{T}$ and $\mathcal{R}$ has nonzero probability for every θ. The same is true of $\mathcal{S}$ although $\mathcal{S}$ is not necessarily the direct product of $\mathcal{T}$ and $\mathcal{R}$. For any point (t, r) in $\mathcal{S}$ both $P_\theta(T=t|R=r)$ and $P_\theta(R=r|T=t)$ are defined and nonzero for every θ. By sufficiency of T, the latter is the same for all θ and therefore can be denoted by $C(r, t)$. Now consider any fixed r in $\mathcal{R}$. We can write

$$(10)\quad \sum_t f(t)P_\theta(T=t|R=r) = \frac{1}{P_\theta(R=r)} \sum_t f(t)C(r, t)P_\theta(T=t).$$

If we assume (6) holds, then the above expression equals zero for all θ. The assumed completeness of $T(x)$ then implies $f(t)C(r, t) = 0$ for all t in $\mathcal{T}$. Since $C(r, t) \neq 0$ for (r, t) in $\mathcal{S}$, we have $f(t) = 0$ whenever (r, t) in $\mathcal{S}$, that is, whenever $P_\theta(T=t|R=r) > 0$, so that (7) holds.

We now state without proof the modification of the standard Rao-Blackwell Theory.

Theorem 4. Let $\hat{\varphi}$ be any estimator of $\emptyset$ such that $E_\theta(\hat{\varphi}|R = r) = \varphi(r, \theta)$. Let T be conditionally sufficient for θ given R. Define $\tilde{\varphi} = E(\hat{\varphi}|R, T)$. Then $\tilde{\varphi}$ is also a conditionally unbiased estimator of φ given R, and $\mathrm{Var}(\tilde{\varphi}|R) \leq \mathrm{Var}(\hat{\varphi}|R)$.

Lemma 1. Let $\hat{\varphi}(x)$ and $\hat{\Psi}(x)$ be two conditionally unbiased estimators of $\varphi(R, \theta)$ based on a conditionally complete sufficient statistic. Then $\hat{\varphi}(x) = \hat{\Psi}(x)$ a.e. $P_\theta^{X|R}$.

Theorem 5. Let T be conditionally complete sufficient for θ. Let $\varphi(R, \theta)$ be any quantity for which a conditionally unbiased estimator given R exists. Then $\varphi(R, \theta)$ has a unique (a.e.) conditionally UMVU estimator which is a function of T and R.

Thus when a conditionally unbiased estimator is known, we can find a conditionally UMVU estimator by calculating the conditional expectation given both R and T.

4. EXAMPLES

4.1 Example 1. Let $x_1, \ldots, x_n$ be a sample from $N(\mu, 1)$ and let $R(x) = \Sigma c_i x_i$. To avoid a degenerate case we assume not all c_i are equal. If $y_j = x_j$, $j = 1, \ldots, n-1$, and $y_n = R(x)$, then the conditional distribution of $y_1, \ldots, y_{n-1}$ given y_n is multinormal with mean

$$(11) \quad \underset{\sim}{\nu} = d^{-1} r \underset{\sim}{c} + \mu(\underset{\sim}{1} - d^{-1} b \underset{\sim}{c})$$

and covariance matrix

$$(12) \quad C = I_{n-1} - d^{-1} \underset{\sim}{c}\underset{\sim}{c}',$$

where

(13) $b = \sum_1^n c_i$, $d = \sum_1^n c_i^2$, $\underset{\sim}{c} = (c_1, \ldots, c_{n-1})'$, $\underset{\sim}{1} = (1, \ldots, 1)'$.

We find

(14) $\partial \log g_\mu(y_1, \ldots, y_{n-1}|r)/\partial\mu = (\underset{\sim}{1} - d^{-1}b\underset{\sim}{c})'C^{-1}(\underset{\sim}{y} - \underset{\sim}{\nu})$

and

(15) $E(S^2|r) = V(\frac{\partial\varphi}{\partial\mu})^{-2}$, $V = (\underset{\sim}{1} - d^{-1}b\underset{\sim}{c})'C^{-1}(\underset{\sim}{1} - d^{-1}b\underset{\sim}{c})$.

If we specialize to $n = 3$, $R(x) = x_1 + x_2 + 2x_3$, then any conditionally unbiased estimator of $\varphi(R, \mu)$ satisfies

$$\mathrm{Var}(\hat{\varphi}|r) \geq 3(\partial\varphi/\partial\mu)^2.$$

This bound may or may not be achievable depending on the form of $\varphi(R, \mu)$. It is, for example, achieved when $\varphi(R, \mu) = R\mu$ if we take $\hat{\varphi} = R\hat{\mu}$ with $\hat{\mu} = \frac{3}{2}(x_1 + x_2) - \frac{1}{2}r = x_1 + x_2 - x_3$. The choice $\hat{\mu} = \bar{x}$ would not give a conditionally unbiased estimator.

4.2 Example 2.

4.2.1 A problem in classification theory. The present paper was in fact motivated by some problems in classification theory which are more fully discussed in [7]. Let S_1 and S_2 denote samples of size N_1 and N_2 known to come from populations π_1 and π_2 respectively. The samples determine a rule for classifying a future observation as either belonging to π_1 or to π_2. We wish to estimate the probabilities of misclassifying an observation from π_1 in π_2 or vice versa. These two probabilities depend on the samples and on unknown population parameters.

Consider next the p-variate normal case with unknown mean vectors $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ and known and equal covariance matrices Σ. If

$\bar{\underset{\sim}{x}}_1$, $\bar{\underset{\sim}{x}}_2$ are the sample mean vectors and $\underset{\sim}{z}$ is a future observation, then the usual symmetrical classification rule is (Anderson [1])

(16) classify $\underset{\sim}{z}$ as π_1 if $(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)'\Sigma^{-1}\underset{\sim}{z} \geq \frac{1}{2}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)'\Sigma^{-1}(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2)$ and otherwise classify $\underset{\sim}{z}$ as π_2. For definiteness consider the probability P_2 of misclassifying an observation from π_2 into π_1 for given values of $\bar{\underset{\sim}{x}}_1$, $\bar{\underset{\sim}{x}}_2$, which is

(17) $P_2 = 1 - F(C)$

where

(17a) $C = \frac{1}{2}D + D^{-1}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)\Sigma^{-1}(\bar{\underset{\sim}{x}}_2 - \underset{\sim}{\mu}_2)$,

where D is the positive square root of

$$D^2 = (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)'\Sigma^{-1}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2),$$

and where F is the standard normal C.D.F. P_2 is a function of $\bar{\underset{\sim}{x}}_1$, $\bar{\underset{\sim}{x}}_2$ and $\underset{\sim}{\mu}_2$ which we wish to estimate. In this case if we attempt to estimate conditionally on $(\bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2)$, then there is no conditionally unbiased estimator of P_2 because P_2 depends on $\underset{\sim}{\mu}_2$ but the conditional distribution of (S_1, S_2) given $(\bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2)$ is the same for all $(\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2)$, owing to the sufficiency of $(\bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2)$.

The theory of conditional unbiased estimation will however apply in the case where the classification rule is tested by using it on additional observations of known origin. Although this does not seem to be a widely used method, we find it mentioned, for example by Hills [4] and Lachenbruch and Mickey [5]. Let $t_1, \ldots, t_m$ be additional observations from π_2 and let q be the proportion of these m observations which are misclassified.

Then q is clearly a conditionally unbiased estimator of P_2 since its conditional distribution given $\underset{\sim}{\bar{x}}_1$, $\underset{\sim}{\bar{x}}_2$ is binomial with mean equal to P_2. (The argument does not involve the normality assumption, so that the method is clearly quite general).

4.2.2 The Cramér-Rao bound. To illustrate Theorem 1 we will apply it to the problem just described. Since Theorem 1 applies only for scalar parameters we specialize Section 4.2.1 to the univariate normal case where π_1 and π_2 are $N(\mu_1, 1)$ and $N(\mu_2, 1)$. The function φ of Section 2 corresponds to P_2 given by (17), and in the univariate case (17) reduces to

$$(18)\quad P_2 = \begin{cases} 1-F(c) & \text{if } \bar{x}_1 > \bar{x}_2 \\ F(c) & \text{if } \bar{x}_1 < \bar{x}_2 \end{cases}$$

where

$$(19)\quad c = \frac{1}{2}\bar{x}_1 + \frac{1}{2}\bar{x}_2 - \mu_2.$$

The conditioning variate y_n of Theorem 1 corresponds to the two sample means $\bar{x}_1$, $\bar{x}_2$. For $y_1, \ldots, y_{n-1}$ we may choose $N_1 + N_2 + m - 2$ suitable coordinates such that, together with $\bar{x}_1$, $\bar{x}_2$, these coordinates are in one-to-one correspondence with the sample $(x_{11}, \ldots, x_{1N_1}, x_{21}, \ldots, x_{2N_2}, t_1, \ldots, t_m)$. Our arbitrary choice is to delete x_{1N_1} and x_{2N_2} from the above array to obtain $y_1, \ldots, y_{n-1}$. The desired conditional distribution is found to have mean vector $(\bar{x}_1, \ldots, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_2, \mu_2, \ldots, \mu_2)$ where the blocks are of length N_1-1, N_2-1 and m respectively. Since μ_2 appears only in the last m positions and since the corresponding variates are (unconditionally and conditionally) independent of the others we find

$$(20)\quad \partial \log g_{\mu_2} / \partial \mu_2 = \sum_1^m (t_i - \mu_2).$$

It follows that

$$(21)\quad E(S^2|\bar{x}_1, \bar{x}_2) = (\frac{\partial\varphi}{\partial\mu_2})^{-2}E\{\sum_{i=1}^{m}(t_i - \mu_2)^2|\bar{x}_1, \bar{x}_2\}$$

$$= m(\frac{\partial\varphi}{\partial\mu_2})^{-2}.$$

Whether $\bar{x}_1$ be greater than or less than $\bar{x}_2$ we find

$$(22)\quad (\frac{\partial\varphi}{\partial\mu_2})^2 = f^2(c)$$

where f is the standard normal density function. Thus the Cramer-Rao bound is

$$(23)\quad Var(\hat{\varphi}|\bar{x}_1, \bar{x}_2) \geq \frac{1}{m} f^2(c).$$

4.2.3 The conditional UMVU estimator. The estimator q of Section 4.2.1 can be improved with respect to its conditional variance by the conditional Rao-Blackwell method. Let $\underset{\sim}{\bar{t}} = \frac{1}{m}\Sigma\, \underset{\sim}{t}_i$. Then $\underset{\sim}{\bar{t}}$ is conditionally complete for $\underset{\sim}{\mu}_2$ given $\underset{\sim}{\bar{x}}_1$, $\underset{\sim}{\bar{x}}_2$. The conditional expectation of any unbiased estimator of P_2 will lead to the conditional UMVU estimator, and for convenience we take $\hat{\varphi} = 1$ or 0 according as $\underset{\sim}{t}_1$ is misclassified as not by the rule (16). The UMVU estimator $\tilde{\varphi}$ is then just the conditional probability

$$(24)\quad \tilde{\varphi} = P\{(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)\Sigma^{-1}\underset{\sim}{t}_1 > \frac{1}{2}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)'\Sigma^{-1}(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2)|\underset{\sim}{\bar{x}}_1, \underset{\sim}{\bar{x}}_2, \underset{\sim}{\bar{t}}\}.$$

To evaluate this we first find the conditional distribution of $\underset{\sim}{t}_1$ given $\underset{\sim}{\bar{t}}$ to be $N_p(\underset{\sim}{\bar{t}}, \frac{m-1}{m}\Sigma)$. It follows that

$$(25)\quad \tilde{\varphi} = 1\text{-}F((1\text{-}m^{-1})^{-1/2}[\tfrac{1}{2}D + D^{-1}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)'\Sigma^{-1}(\underset{\sim}{\bar{x}}_2 - \underset{\sim}{\bar{t}})]).$$

It is reasonably straightforward to get the conditional variance of $\tilde{\varphi}$ in terms of the bivariate normal distribution function. We find

$$(26)\quad \mathrm{Var}(\tilde{\varphi}|\bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2) = F(C, C; m^{-1}) - F(C, C, 0)$$

where C is defined by (17a) and where $F(a, b; \rho) = P(u \le a, v \le b)$ with (u, v) joint normal with zero means, unit variances and correlation ρ.

In the univariate case, (26) holds with C replaced by c defined in (19). Table 1 compares the variance of $\tilde{\varphi}$ with that of q and with the Cramér-Rao bound for several values of c and m.

TABLE 1

m	c	Var q	Var $\tilde{\varphi}$	C-R Bound
5	0.0	0.0500	0.0320	0.0318
	0.5	0.0427	0.0255	0.0248
	1.0	0.0267	0.0129	0.0117
	1.5	0.0125	0.0041	0.0034
	2.0	0.0044	0.0009	0.0006
10	0.0	0.0250	0.0159	0.0159
	0.5	0.0213	0.0126	0.0124
	1.0	0.0134	0.0061	0.0059
	1.5	0.0062	0.0018	0.0017
	2.0	0.0022	0.0004	0.0003
20	0.0	0.0125	0.0080	0.0080
	0.5	0.0107	0.0062	0.0062
	1.0	0.0067	0.0030	0.0029
	1.5	0.0031	0.0009	0.0008
	2.0	0.0011	0.0002	0.0001

REFERENCES

[1] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1958.

[2] Bahadur, R. R., "Sufficiency and Statistical Decision Functions," Annals of Mathematical Statistics, Vol. 25 (1954), pp. 423-462.

[3] Chapman, D. G. and Robbins, H., "Minimum Variance Estimation Without Regularity Assumptions," Annals of Mathematical Statistics, Vol. 22 (1951), pp. 581-586.

[4] Hills, M., "Allocation Rules and Their Error Rates," Journal of the Royal Statistical Society, Series B, Vol. 28 (1966), pp. 1-32.

[5] Lachenbruch, P. A. and Mickey, M. R., "Estimation of Error Rates in Discriminant Analysis," Technometrics, Vol. 10 (1968), pp. 715-725.

[6] Lehmann, E. L., Notes on the Theory of Estimation, Statistics Dept., Univ. of California, Berkeley, 1950.

[7] Sorum, Marilyn J., "Estimating the Probability of Misclassification," Technical Report No. 110 (1968), Statistics Dept., Univ. of Minnesota, Minneapolis.