Biometrika (1972), 59, 1, *p.* 19
With 1 *text-figure*
Printed in Great Britain

Separate sample logistic discrimination

BY J. A. ANDERSON
University of Oxford

SUMMARY

The problem of discrimination when all or most of the observations are qualitative is discussed. The method of logistic discrimination introduced by Cox (1966) and Day & Kerridge (1967) is extended to the situation where separate samples are taken from each population, using the results of Aitchison & Silvey (1958) on constrained maximum likelihood estimation. The method is further extended to discrimination between three or more populations. The properties of logistic discrimination are investigated by simulation and the method is applied to the differential diagnosis of *kerato-conjunctivitis sicca.*

Some key words: Discrimination with qualitative observations; Logistic discrimination; Constrained maximum likelihood; Simulation study; Application to medical diagnosis.

1. INTRODUCTION

Cox (1966) and Day & Kerridge (1967) both suggested the logistic form for posterior probabilities as a basis for discrimination between two populations. Day and Kerridge specifically considered estimating the allocation rule when sampling was from the mixture of populations, while Cox was not concerned with discriminant estimation. However, a discriminant function is often required when samples are taken from each population separately, perhaps to allocate further sample points from the mixture of populations in known or estimable proportions. In this case the allocation rule would be based on posterior probabilities. Alternatively, if the proportions cannot be specified, the rule would depend on likelihood ratios. It is intended to extend the application of logistic discrimination to these two situations and to more than two populations.

The techniques of this paper were developed for use with data that include polychotomous observations. In particular it is thought that applications to medical diagnosis are relevant and an example of this is given in §6.

2. FORMULATION OF THE PROBLEM

Suppose that sample points $\mathbf{x}^T = (x_1, \ldots, x_p)$ are available from the population H_s and that the likelihood of $\mathbf{x}$ given H_s is $f_s(\mathbf{x})$ $(s = 1, \ldots, k)$. All the components of $\mathbf{x}$ are real but some are continuous and some are polychotomous. The discrimination problem is to find a rule for allocating further points $\mathbf{x}$ of unknown origin to populations.

If it is known that the points to be allocated are from a mixture of the distributions $H_1, \ldots, H_k$ in the proportion $\mathbf{\Pi}^T = (\Pi_1, \ldots, \Pi_k)$, where

$$\sum_{s=1}^{k} \Pi_s = 1,$$

then the simplest optimizing method of discrimination is to maximize the probability of correct allocation. This is achieved by allocating the sample point $\mathbf{x}$ to H_s if

$$\Pi_s f_s(\mathbf{x}) \geqslant \Pi_t f_t(\mathbf{x}) \quad (t = 1, \ldots, k;\ t \neq s), \tag{1}$$

or

$$\mathrm{pr}\,(H_s|\mathbf{x}) \geqslant \mathrm{pr}\,(H_t|\mathbf{x}) \quad (t = 1, \ldots, k;\ t \neq s) \tag{2}$$

(Rao, 1965). Attention will be directed towards this allocation rule because it is the simplest and perhaps the most general. However, the method developed in this paper can be used to advantage with other optimizing techniques. This point is discussed in §7.

The Cox–Day–Kerridge approach was to postulate the logistic form for the posterior probabilities when $k = 2$;

$$\mathrm{pr}\,(H_1|\mathbf{x}) = \exp\,(\alpha_0 + \alpha_1 x_1 + \ldots + \alpha_p x_p)\,\mathrm{pr}\,(H_2|\mathbf{x}), \tag{3}$$

$$\mathrm{pr}\,(H_2|\mathbf{x}) = 1/\{1 + \exp\,(\alpha_0 + \alpha_1 x_1 + \ldots + \alpha_p x_p)\}. \tag{4}$$

There is an obvious extension of this to k populations (Cox, 1970, p. 104):

$$\mathrm{pr}\,(H_s|\mathbf{x}) = p_{s\mathbf{x}} = \exp\{(1, \mathbf{x}^T)\,\boldsymbol{\alpha}_s\}\,\mathrm{pr}\,(H_k|\mathbf{x}),$$

$$\mathrm{pr}\,(H_k|\mathbf{x}) = p_{k\mathbf{x}} = 1\Big/\left[1 + \sum_{s=1}^{k-1} \exp\{(1, \mathbf{x}^T)\,\boldsymbol{\alpha}_s\}\right], \tag{5}$$

where $\boldsymbol{\alpha}_s^T = (\alpha_{s0,}\alpha_{s1}, \ldots, \alpha_{sp})$ $(s = 1,, \ldots k-1)$. The next step of their procedure was to estimate the $\boldsymbol{\alpha}_s$ directly from sample points of the mixture. Only the parameters that are specifically required for the discrimination rule are estimated and, further, the method is exactly the same for all $f_s(\mathbf{x})$ satisfying (5). By contrast, the classical approach is to derive the allocation rule from estimates of all the parameters postulated in each distribution with the result that, first, many more parameters have to be estimated and, secondly, different methods are required for different assumed families of distributions $\{f_s(\mathbf{x})\}$. The advantages of the Cox–Day–Kerridge approach are clearer when it is realized that equation (5) is satisfied by many of the families commonly postulated in discrimination. These include:

(i) multivariate normal with equal dispersion matrices;

(ii) multivariate independent dichotomous, 0 or 1, variables;

(iii) multivariate dichotomous variates following the log linear model (Birch, 1963) with second and higher order effects the same in each population;

(iv) a combination of (i) and (iii).

Model (5) can be given even greater generality by including extra terms to allow for different second order interactions; for example, different dispersion matrices in (i). However, this is at the cost of introducing more parameters which all have to be estimated. It is clear, then, that (5) will be true or approximately true, in a number of practical applications.

The Cox–Day–Kerridge formulation was to assume that sample points $\mathbf{x}$ were available from the mixture of two populations in unknown proportions. Generalizing this to k populations, suppose that $n_{s\mathbf{x}}$ sample points are observed from H_s at the point $\mathbf{x}$ $(s = 1, \ldots, k)$. Usually most of the $n_{s\mathbf{x}}$ will be zero and the rest unity. Let $n_s = \Sigma n_{s\mathbf{x}}$. Thus n_s $(s = 1, \ldots, k)$ is the total sample from H_s and is a random variable. The likelihood of the observations is

$$L \propto \prod_{\mathbf{x}} \prod_{s=1}^{k} (p_{s\mathbf{x}}\phi_{\mathbf{x}})^{n_{s\mathbf{x}}},$$

where $\phi_{\mathbf{x}}$ is the probability or probability density of the mixture distribution at $\mathbf{x}$. It follows, after some reduction, that the maximum likelihood equations for the matrix of coefficients, $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{k-1}]$, are

$$\frac{\partial \log L}{\partial \alpha_{sj}} = \sum_{\mathbf{x}} (n_{s\mathbf{x}} - n_{\mathbf{x}} p_{s\mathbf{x}}) x_j = 0 \quad (s = 1, \ldots, k-1; j = 0, \ldots, p), \tag{6}$$

where $n_{\mathbf{x}} = \Sigma n_{s\mathbf{x}}$ for all $\mathbf{x}$. The solution of these equations is dealt with in some detail for $k = 2$ by Cox (1970, p. 87).

The problem to be considered in this paper is to estimate logistic posterior probabilities of the type (5) but now where samples are available from each population separately, so the n_s are fixed. To define the mixture of populations to which the posterior probabilities refer, it is necessary to specify the mixing proportions

$$\boldsymbol{\Pi}^T = (\Pi_1, \ldots, \Pi_k), \quad \sum_{s=1}^{k} \Pi_s = 1,$$

of $H_1, \ldots, H_k$. These can be either given or estimated from other data. The situation where this is not possible is discussed in §3·4, using likelihood ratios.

With the above notation, the joint likelihood of the separate samples from $H_1, \ldots, H_k$ can be written

$$L \propto \prod_{s=1}^{k} \prod_{\mathbf{x}} \{L(\mathbf{x}|H_s)\}^{n_{s\mathbf{x}}}. \tag{7}$$

Suppose further that x_j $(j = 0, 1, \ldots, p)$ is dichotomous, with values 0 or 1, enabling likelihoods to be regarded as probabilities. This condition will be relaxed in §3·3. Then,

$$L(\mathbf{x}|H_s) = \mathrm{pr}(\mathbf{x}|H_s) = \frac{\mathrm{pr}(H_s|\mathbf{x})\,\mathrm{pr}(\mathbf{x})}{\mathrm{pr}(H_s)}.$$

As before, let $\phi_{\mathbf{x}} = \mathrm{pr}(\mathbf{x})$; then

$$L(\mathbf{x}|H_s) = p_{s\mathbf{x}} \phi_{\mathbf{x}} / \Pi_s \tag{8}$$

and when substituted into (7) this gives

$$\log L = \mathrm{const} + \sum_{s=1}^{k} \sum_{\mathbf{x}} \{n_{s\mathbf{x}} \log (p_{s\mathbf{x}} \phi_{\mathbf{x}})\}, \tag{9}$$

where the constant includes all terms independent of the $\boldsymbol{\alpha}_s$'s and $\phi_{\mathbf{x}}$'s. This likelihood is the same as that of the Cox–Day–Kerridge formulation but it must be remembered that $\boldsymbol{\alpha}_s$'s and $\phi_{\mathbf{x}}$'s are related by the functionally independent conditions

$$\sum_{\mathbf{x}} \phi_{\mathbf{x}} = 1, \quad \sum_{\mathbf{x}} p_{s\mathbf{x}} \phi_{\mathbf{x}} = \Pi_s \quad (s = 1, \ldots, k-1), \tag{10}$$

where the summations are over all possible $\mathbf{x}$ values. Thus the problems of estimating $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{k-1})$ in the two sampling situations are not the same.

The estimation of $\mathbf{A}$ when the log likelihood is given by (9) subject to the conditions (10) is the central problem of the paper. A fairly obvious approach is to take as estimators those values of the parameters that maximize the likelihood subject to the constraints. Aitchison & Silvey (1958) showed that this method of constrained maximum likelihood has similar properties to ordinary maximum likelihood estimation.

The expression (9) for $\log L$ can be maximized subject to the conditions (10) quite straightforwardly but the algebra is heavy and the asymptotic dispersion matrix $\mathbf{D}$ has a very unwieldy form. However, an easier solution exists with a much more manageable expression for $\mathbf{D}$. This will be given in the next section.

3. The maximum likelihood estimation of separate sample logistic discriminators: theoretical considerations

It is required to maximize $\log L$ given by (9) subject to the constraints (10). It must be remembered that the samples are taken from $L(\mathbf{x}|H_s)$ $(s = 1, \ldots, k)$ and that $\mathbf{\Pi}$ is fixed separately. For the same set of k samples different logistic functions (5) can be estimated corresponding to different values of $\mathbf{\Pi}$; moreover, there is a very simple relationship between them.

Suppose then that there is one set of proportions $\mathbf{\Pi}$ to which all the foregoing equations refer and another set $\mathbf{\Pi}'$, so that

$$\begin{aligned} \mathrm{pr}'(H_s|\mathbf{x}) &= p'_{s\mathbf{x}} \quad (s = 1, \ldots, k), \\ \mathrm{pr}'(\mathbf{x}) &= \phi'_{\mathbf{x}} \quad \text{all } \mathbf{x}. \end{aligned} \tag{11}$$

The samples from $L(\mathbf{x}|H_s)$ do not depend on the choice of $\mathbf{\Pi}$ or $\mathbf{\Pi}'$, so

$$L(\mathbf{x}|H_s) = p_{s\mathbf{x}}\phi_{\mathbf{x}}/\Pi_s = p'_{s\mathbf{x}}\phi'_{\mathbf{x}}/\Pi'_s \quad (s = 1, \ldots, k). \tag{12}$$

Dividing through these equations by the kth, we find that

$$\frac{p'_{s\mathbf{x}}}{p'_{k\mathbf{x}}} = \frac{p_{s\mathbf{x}}}{p_{k\mathbf{x}}} \frac{\Pi_k \Pi'_s}{\Pi'_k \Pi_s} \quad (s = 1, \ldots, k-1) \tag{13}$$

or

$$p'_{s\mathbf{x}}/p'_{k\mathbf{x}} = \exp\{\beta_s + (1, \mathbf{x}^T)\boldsymbol{\alpha}_s\} \quad (s = 1, \ldots, k-1), \tag{14}$$

where $\beta_s = \log(\Pi_k \Pi'_s) - \log(\Pi'_k \Pi_s)$. Using $\Sigma p'_{s\mathbf{x}} = 1$, we obtain

$$\begin{aligned} p'_{s\mathbf{x}} &= \exp\{(1, \mathbf{x}^T)\boldsymbol{\alpha}'_s\} p'_{k\mathbf{x}} \quad (s = 1, \ldots, k-1), \\ p'_{k\mathbf{x}} &= 1\Big/\left[1 + \sum_{s=1}^{k-1} \exp\{(1, \mathbf{x}^T)\boldsymbol{\alpha}'_s\}\right], \end{aligned} \tag{15}$$

where

$$\alpha'_{s0} = \alpha_{s0} + \beta_s, \quad \alpha'_{sj} = \alpha_{sj} \quad (j = 1, \ldots, p). \tag{16}$$

Note that this does not imply that $\phi_{\mathbf{x}} = \phi'_{\mathbf{x}}$. From (15) and (16), it can be seen that the maximum likelihood estimates of the parameters $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{k-1}]$ with $\mathbf{\Pi}$ can be obtained directly from those of $\mathbf{A}' = [\boldsymbol{\alpha}'_1, \ldots, \boldsymbol{\alpha}'_{k-1}]$ with $\mathbf{\Pi}'$, and vice versa.

In particular, it became apparent when working on the maximization of the likelihood function (9) subject to the conditions (10) that a simplification would be achieved by estimating the logistic functions appropriate to the proportions $\Pi^*_s = n_s/n$ $(s = 1, \ldots, k)$, where $n = \Sigma n_s$, and deriving the logistic functions for any other $\mathbf{\Pi}$ subsequently.

3·1. *The maximum likelihood equations when* $\mathbf{\Pi} = \mathbf{\Pi}^*$

Lagrange multipliers will now be used to maximize the likelihood function (9) subject to the conditions (10) with $\mathbf{\Pi} = \mathbf{\Pi}^*$. For ease of notation, x_0 is defined to be 1 at each sample point. Then

$$\begin{aligned} \frac{\partial p_{s\mathbf{x}}}{\partial \alpha_{sj}} &= p_{s\mathbf{x}}(1 - p_{s\mathbf{x}})x_j \quad (s = 1, \ldots, k-1; j = 0, 1, \ldots, p), \\ \frac{\partial p_{s\mathbf{x}}}{\partial \alpha_{tj}} &= -p_{s\mathbf{x}} p_{t\mathbf{x}} x_j \quad (s, t = 1, \ldots, k-1, s \neq t; j = 0, 1, \ldots, p), \end{aligned} \tag{17}$$

and it follows after some reduction that

$$\frac{\partial \log L}{\partial \alpha_{sj}} = \sum_{\mathbf{x}} (n_{s\mathbf{x}} - n_{\mathbf{x}} p_{s\mathbf{x}})\, x_j,$$

where $n_{\mathbf{x}} = \Sigma n_{s\mathbf{x}}$, for all $\mathbf{x}$; $\partial \log L/\partial \phi_{\mathbf{x}} = n_{\mathbf{x}}/\phi_{\mathbf{x}}$, so that the equations that give a stationary point are

$$\sum_{\mathbf{x}} (n_{s\mathbf{x}} - n_{\mathbf{x}} p_{s\mathbf{x}})\, x_j + \sum_{t=1}^{k-1} \lambda_t \sum_{\mathbf{x}} (-p_{s\mathbf{x}} p_{t\mathbf{x}} \phi_{\mathbf{x}} x_j) + \lambda_s \sum_{\mathbf{x}} p_{s\mathbf{x}}(1 - p_{s\mathbf{x}})\, \phi_{\mathbf{x}} x_j = 0$$

$$(s = 1, \ldots, k-1;\; j = 0, 1, \ldots, p), \tag{18}$$

$$n_{\mathbf{x}}/\phi_{\mathbf{x}} + \mu + \sum_{t=1}^{k-1} \lambda_t p_{t\mathbf{x}} = 0 \quad \text{(for all x)} \tag{19}$$

and equations (10), where $\mu, \lambda_1, \ldots, \lambda_{k-1}$ are the undetermined multipliers corresponding, in order, to the k conditions in (10). Equations (18) and (19) can be written as

$$\sum_{\mathbf{x}} \left(n_{s\mathbf{x}} - n_{\mathbf{x}} p_{s\mathbf{x}} + \lambda_s p_{s\mathbf{x}} \phi_{\mathbf{x}} - p_{s\mathbf{x}} \sum_{t=1}^{k-1} \lambda_t p_{t\mathbf{x}} \phi_{\mathbf{x}} \right) x_j = 0 \quad (s = 1, \ldots, k-1; j = 0, 1, \ldots, p), \tag{20}$$

$$n_{\mathbf{x}} + \mu \phi_{\mathbf{x}} + \sum_{t=1}^{k-1} \lambda_t p_{t\mathbf{x}} \phi_{\mathbf{x}} = 0 \quad \text{(for all x)}. \tag{21}$$

Substituting $\sum_{t=1}^{k-1} \lambda_t p_{t\mathbf{x}} \phi_{\mathbf{x}}$ from (21) into (20), we obtain

$$\sum_{\mathbf{x}} \{ n_{s\mathbf{x}} - n_{\mathbf{x}} p_{s\mathbf{x}} + \lambda_s p_{s\mathbf{x}} \phi_{\mathbf{x}} + p_{s\mathbf{x}} (n_{\mathbf{x}} + \mu \phi_{\mathbf{x}}) \} x_j = 0,$$

$$\sum_{\mathbf{x}} \{ n_{s\mathbf{x}} + p_{s\mathbf{x}} \phi_{\mathbf{x}} (\mu + \lambda_s) \} x_j = 0 \quad (s = 1, \ldots, k-1; j = 0, 1, \ldots, p). \tag{22}$$

In this equation, choose $j = 0$ and $x_0 = 1$ for all $\mathbf{x}$, so that

$$\sum_{\mathbf{x}} \{ n_{s\mathbf{x}} + p_{s\mathbf{x}} \phi_{\mathbf{x}} (\mu + \lambda_s) \} = 0$$

or

$$n_s + \Pi_s^* (\mu + \lambda_s) = 0, \quad \mu + \lambda_s = -n \quad (s = 1, \ldots, k-1), \tag{23}$$

since $\Pi_s^* = n_s/n$. Summing (21) over all $\mathbf{x}$ gives

$$n + \mu + \sum_{t=1}^{k-1} \lambda_t \Pi_t^* = 0$$

or

$$n + \mu + (1/n) \sum_{t=1}^{k-1} \lambda_t n_t = 0.$$

Together with (23), this implies that $\mu = -n$ and $\lambda_t = 0$ for $t = 1, \ldots, k-1$. Thus (21) becomes

$$\phi_{\mathbf{x}} = n_{\mathbf{x}}/n \tag{24}$$

and (22) becomes

$$\sum_{\mathbf{x}} (n_{s\mathbf{x}} - n_{\mathbf{x}} p_{s\mathbf{x}})\, x_j = 0 \quad (s = 1, \ldots, k-1; j = 0, 1, \ldots, p). \tag{25}$$

There are $(k-1)(p+1)$ equations in (25), none of which involves any of the $\phi_{\mathbf{x}}$'s. Thus the large number of nuisance parameters that were included in the likelihood function (9) have now been eliminated, resulting in equations (25) which are formally identical to the maximum likelihood equations (6) for the logistic parameters in the Cox–Day–Kerridge formulation. There is thus a close relationship between the two estimation systems. Indeed, the use of equations (6) for estimating $\mathbf{A}$ with fixed (n_s) first suggested itself to the author as an approximation, the idea being to adjust (α_{s0}) afterwards to give the desired proportions. However, it can be seen that this is a constrained maximum likelihood solution with corresponding desirable properties (Aitchison & Silvey, 1958).

3·2. *The asymptotic dispersion matrix of the estimators*

Although the properties of constrained and ordinary maximum likelihood estimators are very similar, the forms of the asymptotic dispersion matrices are quite different, since the constrained matrix has to take account of the functional dependence between the parameters. Although the general result has a fearsome appearance, drastic simplifications are obtained when $\mathbf{\Pi} = \mathbf{\Pi}^*$. For the moment suppose that the parameters to be estimated are $\boldsymbol{\theta}^T = (\theta_1, \ldots, \theta_w)$ with k conditions $h_s(\theta) = 0$ $(s = 1, \ldots, k)$. The $w \times w$ matrix

$$\mathbf{B} = E\left(\frac{\partial \log L}{\partial \theta_i}\frac{\partial \log L}{\partial \theta_j}\right) = E\left(-\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}\right),$$

here. The $w \times k$ matrix $\mathbf{H} = \{\partial h_s/\partial \theta_j\}$. Then Aitchison & Silvey's (1958) result for the $w \times w$ asymptotic dispersion matrix $\mathbf{D}$ of the constrained maximum likelihood estimators of $\boldsymbol{\theta}$ is

$$\mathbf{D} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{B}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{B}^{-1}. \qquad (26)$$

There are $(k-1)(p+1)$ α-parameters to estimate and, say, a $\phi_{\mathbf{x}}$-type parameters so that $\mathbf{D}$ can be partitioned as

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{\alpha\alpha} & \mathbf{D}_{\alpha\phi} \\ \mathbf{D}_{\phi\alpha} & \mathbf{D}_{\phi\phi} \end{bmatrix},$$

where, for example $\mathbf{D}_{\alpha\alpha}$ is the $(k-1)(p+1) \times (k-1)(p+1)$ dispersion matrix of the α_s's, which is the only part of $\mathbf{D}$ that is of interest here. The matrix $\mathbf{B}$ can be partitioned in exactly the same way as $\mathbf{D}$. The order in which the parameters are to be taken is $\alpha_{10}, \alpha_{11}, \ldots, \alpha_{1p}$; $\ldots, \alpha_{k-1,0}, \alpha_{k-1,1}, \ldots, \alpha_{k-1,p}$; $\phi_{\mathbf{x}_1}, \ldots, \phi_{\mathbf{x}_a}$, which corresponds to the above partitioning. Suppose that

$$h_k = \sum_{\mathbf{x}} \phi_{\mathbf{x}} - 1, \quad h_s = \sum_{\mathbf{x}} p_{s\mathbf{x}}\phi_{\mathbf{x}} - \Pi_{s\mathbf{x}} = 0 \quad (s = 1, \ldots, k-1),$$

then $\mathbf{H}$ can be partitioned as

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{\alpha\alpha} & \mathbf{H}_{\phi\alpha} \\ \mathbf{H}_{\alpha\phi} & \mathbf{H}_{\phi\phi} \end{bmatrix},$$

where, for example, the $(k-1)(p+1) \times (k-1)$ matrix,

$$\mathbf{H}_{\alpha\alpha} = \left\{\frac{\partial h_t}{\partial \alpha_{sj}}\right\} \quad (t = 1, \ldots, k-1).$$

It is easy to show that all the elements in $\mathbf{H}_{\phi\alpha}$ are zero; $\mathbf{H}_{\phi\phi}^T\mathbf{B}_{\phi\phi}^{-1}\mathbf{H}_{\phi\phi} = 1/n$; $\mathbf{H}_{\alpha\phi}^T\mathbf{B}_{\phi\phi}^{-1}\mathbf{H}_{\phi\phi} = (1/n)(\Pi_1^*, \ldots, \Pi_{k-1}^*)^T$; and that the $(k-1) \times (k-1)$ matrix, $\mathbf{H}_{\alpha\phi}^T\mathbf{B}_{\phi\phi}^{-1}\mathbf{H}_{\alpha\phi} = \{\sum_{\mathbf{x}} p_{s\mathbf{x}}p_{t\mathbf{x}}\phi_{\mathbf{x}}\}$. However the matrix $\mathbf{H}_{\alpha\alpha}^T \mathbf{B}_{\alpha\alpha}^{-1} \mathbf{H}_{\alpha\alpha}$ is more difficult until it is realized that the ith column of $\mathbf{H}_{\alpha\alpha}$ is equal to the $\{1+(i-1)(p+1)\}$th row, and column, of $(1/n)\,\mathbf{B}_{\alpha\alpha}$. It follows that

$$\mathbf{H}_{\alpha\alpha}^T \mathbf{B}_{\alpha\alpha}^{-1} \mathbf{H}_{\alpha\alpha} = (1/n)\operatorname{diag}\big(\sum_{\mathbf{x}} p_{1\mathbf{x}}\phi_{\mathbf{x}}, \ldots, \sum_{\mathbf{x}} p_{k-1,\mathbf{x}}\phi_{\mathbf{x}}\big) - (1/n)\big\{\sum_{\mathbf{x}} p_{s\mathbf{x}} p_{t\mathbf{x}} \phi_{\mathbf{x}}\big\}.$$

Then

$$\mathbf{H}^T\mathbf{B}^{-1}\mathbf{H} = (1/n)\begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 \\ \mathbf{Z}_2^T & 1 \end{bmatrix},$$

where $\mathbf{Z}_1 = \operatorname{diag}(\Pi_1^*, \ldots, \Pi_{k-1}^*)$ and $\mathbf{Z}_2^T = (\Pi_1^*, \ldots, \Pi_{k-1}^*)$, and

$$(\mathbf{H}^T\mathbf{B}^{-1}\mathbf{H})^{-1} = n\begin{bmatrix} \mathbf{Y}_1+\mathbf{Y}_2 & \mathbf{Y}_3 \\ \mathbf{Y}_3 & 1/\Pi_k^* \end{bmatrix},$$

where $\mathbf{Y}_1 = \operatorname{diag}(1/\Pi_1^*, \ldots, 1/\Pi_{k-1}^*)$, $\mathbf{Y}_3^T = (-1/\Pi_k^*, \ldots, -1/\Pi_k^*)$ and $\mathbf{Y}_2$ is the $(k-1)\times(k-1)$ matrix with all its elements equal to $1/\Pi_k^*$. It follows that

$$\mathbf{D}_{\alpha\alpha} = \mathbf{B}_{\alpha\alpha}^{-1} - n_k^{-1}\mathbf{E}_{\alpha\alpha},$$

where $\mathbf{E}_{\alpha\alpha}$ is a $(k-1)(p+1)$ square matrix with all elements zero except for terms in the $\{1+(s-1)(p+1)\}$th row: the diagonal element is $1+n_k/n_s$; that in the $\{1+(t-1)(p+1)\}$th column is 1 $(j = 1, \ldots, k-1)$; the other elements are zero again. Thus the only elements where $\mathbf{D}_{\alpha\alpha}$ and $\mathbf{B}_{\alpha\alpha}^{-1}$ differs are concerned with the variances and covariances of the terms $\{\alpha_{s0}\}$.

It follows from the equation (16) that the maximum likelihood estimators of the parameters of the logistic functions for any given choice of $\mathbf{\Pi}$ are immediately obtainable from those with $\mathbf{\Pi}^*$. In fact, all except α_{s0} $(s = 1, \ldots, k-1)$ which varies by an additive constant are the same irrespective of the choice of $\mathbf{\Pi}$. Hence the dispersion matrix of the estimators also remains the same whatever the choice of $\mathbf{\Pi}$.

There is a close relationship between $\mathbf{D}_{\alpha\alpha}$, derived above, and the equivalent matrix $\mathbf{D}_{\alpha\alpha}^{(M)}$ in the Cox–Day–Kerridge model. Let

$$\mathbf{B}_{\alpha\alpha}^{(M)} = E\left\{-\frac{\partial^2 \log L^{(M)}}{\partial\alpha_{sj}\,\partial\alpha_{tj'}}\right\},$$

then $\mathbf{D}_{\alpha\alpha}^{(M)} = \mathbf{B}_{\alpha\alpha}^{-1}$, where $L^{(M)}$ is as given in the discussion of the Cox–Day–Kerridge formulation in § 2. It follows from equation (6) that

$$\frac{\partial^2 \log L^{(M)}}{\partial\alpha_{sj}\,\partial\alpha_{tj'}} = \sum_{\mathbf{x}} n_{\mathbf{x}} p_{s\mathbf{x}} p_{t\mathbf{x}} x_j x_{j'} \quad (s \neq t),$$

$$\frac{\partial^2 \log L^{(M)}}{\partial\alpha_{sj}\partial\alpha_{sj'}} = -\sum_{\mathbf{x}} n_{\mathbf{x}} p_{s\mathbf{x}}(1-p_{s\mathbf{x}})\,x_j x_{j'} \quad (j,j' = 0, 1, \ldots, p).$$

Since $\log L^{(M)}$ and $\log L$ given by (9) are formally equal, these equations are also true for L. However, in the Cox–Day–Kerridge model $E(n_{\mathbf{x}}) = n\phi_{\mathbf{x}}$, while in the situation introduced in this paper, from (12),

$$E(n_{\mathbf{x}}) = \sum_{s=1}^{k} E(n_{s\mathbf{x}}) = \sum_{s=1}^{k} n_s \frac{p_{s\mathbf{x}}\phi_{\mathbf{x}}}{\Pi_s}.$$

If $\mathbf{\Pi} = \mathbf{\Pi}^*$, $n_s/\Pi_s = n$ and $E(n_{\mathbf{x}}) = n\phi_{\mathbf{x}}$. For no other choice of $\mathbf{\Pi}$ is this expression so simple and this is one of the reasons why the expression for $\mathbf{D}_{\alpha\alpha}$ with $\mathbf{\Pi} = \mathbf{\Pi}^*$ is so convenient. Thus by choosing $\mathbf{\Pi} = \mathbf{\Pi}^*$ and $\mathbf{B}_{\alpha\alpha}^{(M)} = \mathbf{B}_{\alpha\alpha}$, it can be seen that $\mathbf{D}_{\alpha\alpha} = \mathbf{D}_{\alpha\alpha}^{(M)} - (1/n_k)\,\mathbf{E}_{\alpha\alpha}$. Thus, it follows from this result and that at the end of § 3·1 that the estimates of the logistic discrimination rule from a given set of data using the Cox–Day–Kerridge formulation and the one introduced here differ only in the estimates of the α_{s0} and their variances and covariances.

3·3. *Continuous and polychotomous observations*

The discussion of the estimation of logistic discriminators has been restricted, so far, to dichotomous variates. Suppose now that the observations are continuous with a mixture density $\phi_{\mathbf{x}}$ at the point $\mathbf{x}$. The expression (9) for $\log L$ is unchanged, but the conditions (10) become

$$\int \phi_{\mathbf{x}} d\mathbf{x} = 1, \quad \int p_{s\mathbf{x}} \phi_{\mathbf{x}} d\mathbf{x} = \Pi_s \quad (s = 1, \ldots, k-1). \tag{26}$$

The maximization of $\log L$ subject to (26) cannot proceed unless assumptions about the functional form of $\phi_{\mathbf{x}}$ are made. In general, this will entail postulating extra parameters $\boldsymbol{\theta}$ for $\phi_{\mathbf{x}}$ and the ensuing maximum likelihood equations will not reduce to two sets for $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{k-1}]$ and $\boldsymbol{\theta}$ separately. This is almost equivalent to the classical approach of §2 with all its drawbacks. However, if one is prepared to make the continuous distribution discrete and estimate $\phi_{\mathbf{x}} \Delta\mathbf{x}$ solution (25) can be used.

Let

$$\psi_{\mathbf{x}} = \phi_{\mathbf{x}} \Delta\mathbf{x} \quad \text{(for all } \mathbf{x}); \tag{27}$$

then the conditions (26) become, approximately,

$$\sum_{\mathbf{x}} \phi_{\mathbf{x}} \Delta\mathbf{x} = 1, \quad \sum_{\mathbf{x}} \phi_{\mathbf{x}} p_{s\mathbf{x}} \Delta\mathbf{x} = \Pi_s \quad (s = 1, \ldots, k-1)$$

or

$$\sum_{\mathbf{x}} \psi_{\mathbf{x}} = 1, \quad \sum_{\mathbf{x}} \psi_{\mathbf{x}} p_{s\mathbf{x}} = \Pi_s \quad (s = 1, \ldots, k-1). \tag{28}$$

Putting the differential elements into the likelihood, we obtain

$$L \propto \prod_{s=1}^{k-1} \prod_{\mathbf{x}} \left(\frac{p_{s\mathbf{x}} \phi_{\mathbf{x}}}{\Pi_s} \Delta\mathbf{x} \right)^{n_{s\mathbf{x}}}$$

$$= \prod_{s=1}^{k-1} \prod_{\mathbf{x}} \left(\frac{p_{s\mathbf{x}} \psi_{\mathbf{x}}}{\Pi_s} \right)^{n_{s\mathbf{x}}}. \tag{29}$$

The problem of maximizing (29) subject to the conditions (28) is then identical to that considered in §3·1. Situations where continuous and dichotomous data both occur, can be dealt with in a similar way, putting in differential elements for all the continuous variables. It should be noted that this method of making continuous variables discrete is not as drastic as it appears because it is not necessary to divide the range of each variable into predetermined classes. In fact, it can be seen in §5 that the above approach gives good estimators of the true discriminant function when $k = 2$ and the underlying distributions are multivariate normal.

It remains to consider observations that belong to one of three or more distinct categories. The most common situation is trichotomous, say $x = 0, 1, 2$. Sometimes these values will correspond to a true linear ordering and (25) can be used without further justification. More often, the alternatives for x will refer to some qualitative observation, say, hair colour: fair, dark or red. Suppose that $k = 2$ and there is one trichotomous observation as above, then a reasonable model is

$$\mathrm{pr}(H_1 | x = 0) = e^{\alpha}/(1 + e^{\alpha}), \quad \mathrm{pr}(H_1 | x = 1) = e^{\alpha+\beta}/(1 + e^{\alpha+\beta}),$$

$$\mathrm{pr}(H_1 | x = 2) = e^{\alpha+\gamma}/(1 + e^{\alpha+\gamma}).$$

This can be put in the framework of model (5) for posterior probabilities by transforming x into two variables y and z, i.e. $x = 0$ if and only if $y = 0$ and $z = 0$; $x = 1$ if and only if $y = 1$ and $z = 0$, and $x = 2$ if and only if $y = 0$ and $z = 1$, and

$$\mathrm{pr}\,(H_1|y,z) = \frac{e^{\alpha+\beta y+\gamma z}}{1+e^{\alpha+\beta y+\gamma z}}.$$

With the assumption of equal second order effects in the log linear models for all k populations (Birch, 1963), this result can be generalized to p observations. Each of those with more than $r \geqslant 2$ levels is replaced by $r-1$ variables analogous to y and z above. Equation (25) can then be used with the transformed data to estimate the logistic discriminators.

3·4. *Estimation of log likelihood ratios*

There are some discrimination situations where posterior probabilities cannot be used. It may be that there is an underlying mixture but $\mathbf{\Pi}$ is not known nor can it be estimated, or the consideration of posterior probabilities may be excluded on logical grounds. Perhaps the obvious criterion for discrimination in this case is the likelihood ratio or its logarithm and it is pleasing that the logistic function approach of § 3·1 can still be used, provided as always with this technique that samples are available from all the populations separately. Cox (1966) would seem to disagree with this approach but perhaps he had not considered the separate sampling scheme.

From equations (5) and (8), assuming a hypothetical mixture in proportions $\mathbf{\Pi}$,

$$R_{st} = \frac{L(\mathbf{x}|H_s)}{L(\mathbf{x}|H_t)} = \frac{p_{s\mathbf{x}}\Pi_s}{p_{t\mathbf{x}}\Pi_t}, \tag{30}$$

so that

$$\begin{aligned} \log R_{st} &= [1,\mathbf{x}^T](\boldsymbol{\alpha}_s-\boldsymbol{\alpha}_t)-r_{st} \quad (s,t = 1,\ldots,k-1) \\ \log R_{sk} &= [1,\mathbf{x}^T]\,\boldsymbol{\alpha}_s-r_{sk} \qquad\quad (s = 1,\ldots,k-1), \end{aligned} \tag{31}$$

where $r_{st} = \log(\Pi_s/\Pi_t)$ $(s,t = 1,\ldots,k)$. Thus the joint likelihood of the separate samples from each population has been reparameterized in terms of the $\boldsymbol{\alpha}$'s, the $\phi_{\mathbf{x}}$'s and an arbitrary set of proportions $\mathbf{\Pi}$, to yield expression (9) for $\log L$. Having estimated the $\boldsymbol{\alpha}$'s using the maximum likelihood equations (25), we find the log likelihood ratios are recovered from (31), using the selected value of $\mathbf{\Pi}$. From what has gone before, there are considerable advantages in choosing $\mathbf{\Pi} = \mathbf{\Pi}^*$.

4. The maximum likelihood estimation of logistic discriminators: practical considerations

4·1. *Iterative solution of the maximum likelihood equations*

Since equations (6) and (25) are identical, there is no point in distinguishing between the two sampling situations in the solution of the maximum likelihood equations. Both Cox (1970, p. 87) and Day & Kerridge (1967) discuss the solution of these equations for $k = 2$. Cox (1970) suggested using least squares on a linear approximation to the logistic function to give starting values for the Newton–Raphson procedure and this could be extended to the case where $k > 2$. However, the method does not work well, as Cox noted, when the samples are well separated. In the author's experience, Newton-Raphson with starting values of zero for all the parameters has usually worked well. In a few cases, one restart from an intermediate set of coefficients has been necessary.

The Newton–Raphson method is easy to apply in this context. Equations (6) and (25) can be written

$$\begin{aligned} f_{sj} &= \sum_{\mathbf{x}} (n_{s\mathbf{x}} - n_{\mathbf{x}} p_{s\mathbf{x}}) x_j = 0 \quad (s = 1, \ldots, k-1; j = 0, 1, \ldots, p), \\ \frac{\partial f_{sj}}{\partial \alpha_{tl}} &= \sum_{\mathbf{x}} n_{\mathbf{x}} p_{s\mathbf{x}} p_{t\mathbf{x}} x_j x_l \quad (s \neq t), \\ \frac{\partial f_{sj}}{\partial \alpha_{sl}} &= - \sum_{\mathbf{x}} n_{\mathbf{x}} p_{s\mathbf{x}} (1 - p_{s\mathbf{x}}) x_j x_l. \end{aligned} \tag{32}$$

Let $\mathbf{F}$ denote the $(k-1)(p+1)$ square matrix with elements

$$F_{sj,tl} = \partial f_{sj} / \partial \alpha_{tl},$$

where the order of the pairs (s,j) and (t,l) is $(1,0)$ $(1,1) \ldots, (1,p)$, $(2,0), \ldots, (2,p), (3,0), \ldots, (3,p), \ldots, (k-1,0), \ldots, (k-1,p)$, and let $\mathbf{f}$ denote the column vector with $(k-1)$ $(p+1)$ rows, $\mathbf{f} = \{f_{sj}\}$ with the same ordering as above. Let $\mathbf{f_a}$ and $\mathbf{F_a}$ denote the values of the above when $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{k-1}$ take on the values indicated by the column vector $\mathbf{a}$, where $\mathbf{a}^T = [\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{k-1}^T]$. Then starting from the set of values $\mathbf{a}_0$, the next value of $\mathbf{a}$ is given by

$$\mathbf{a}_1 = \mathbf{a}_0 - \mathbf{F}_{\mathbf{a}_0}^{-1} \mathbf{f}_{\mathbf{a}_0}. \tag{33}$$

The next value of $\mathbf{a}$, $\mathbf{a}_2$, is obtained from $\mathbf{a}_1$ in the same way and the process is repeated until it converges or until it is clear that convergence will not be reached quickly, if at all, from that particular starting value.

Very often $\mathbf{F}^{-1}$ changes in value slowly in comparison with $\mathbf{f}$ so that it need not be recalculated at each step of the iteration. In the present context, changing $\mathbf{F}^{-1}$ at every tenth step seemed to be satisfactory.

The asymptotic dispersion matrix can be estimated from $\mathbf{F}^{-1}$. It was shown in § 3·2 that $\mathbf{B}_{\alpha\alpha} = \mathbf{B}_{\alpha\alpha}^{(M)}$ for $\boldsymbol{\Pi} = \boldsymbol{\Pi}^*$. So if there is separate sampling from each population, take $\boldsymbol{\Pi} = \boldsymbol{\Pi}^*$, then, in addition, $E(n_{\mathbf{x}}) = n\phi_{\mathbf{x}}$ for both sampling schemes. From equation (24) the maximum likelihood estimate of $\phi_{\mathbf{x}}$ is $n_{\mathbf{x}}/n$ and the value of $\mathbf{B}_{\alpha\alpha}^{-1}$ with the maximum likelihood estimates of the unknown parameters substituted is just the final value of $\mathbf{F}^{-1}$. Hence this is the estimate of the asymptotic dispersion matrix if sampling from the mixture, otherwise the adjustments noted in § 3·2 are required.

From now on the difference between the two sampling schemes will be largely ignored since there is so little difference between the methods of estimation.

4·2 *Complete separation of the sample points*

A major difficulty in the maximum likelihood estimation of $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{k-1}]$ is that in certain situations there is a non-unique maximum of L for infinite values of $\mathbf{A}$. To see this, at each sample point $\mathbf{x}$, which is $[1, x_1, \ldots, x_p]$ since § 3, let

$$z_s = \mathbf{x}^T \boldsymbol{\alpha}_s \quad (s = 1, \ldots, k-1), \quad z_k = 0. \tag{34}$$

The estimate of $\mathbf{A}$, $\mathbf{A}^+ = [\boldsymbol{\alpha}_1^+, \ldots, \boldsymbol{\alpha}_{k-1}^+]$, with corresponding z_s^+ $(s = 1, \ldots, k)$, is said to give complete separation of the sample points if

$$z_s^+ \geqslant z_t^+ \quad (t = 1, \ldots, k; t \neq s) \tag{35}$$

for each $\mathbf{x}$ from H_s $(s = 1, \ldots, k)$. Then with $\mathbf{A} = \mathbf{A}^+$, the likelihood of the n_s sample points from H_s is

$$L_s^+ \propto \prod_{i=1}^{n_s} \frac{\exp(z_{si}^+)}{\sum_{t=1}^{k} \exp(z_{ti}^+)}, \tag{36}$$

where z_{si}^+ is given by (34) with $\mathbf{A} = \mathbf{A}^+$ for the ith point from H_s $(s = 1, \ldots, k)$. If $r\mathbf{A}^+$ is chosen instead of $\mathbf{A}^+$, the likelihood is

$$L_s' \propto \prod_{i=1}^{n_s} \frac{1}{\sum_{t=1}^{k} \exp\{-r(z_{si}^+ - z_{ti}^+)\}}, \tag{37}$$

where $z_{si}^+ - z_{ti}^+ \geqslant 0$, for $t = 1, \ldots, k$, by (35). Hence as $r \to \infty$, all such L_s' terms in the likelihood function tend to 1. Given one $\mathbf{A}^+$ with the property (35), there will generally be others, giving a non-unique maximum of the likelihood function at infinity. In these circumstances, any of the maximum likelihood estimates of $\mathbf{A}$ should give a reasonably good discrimination rule, particularly if all the n_s are large, since all the given sample points are correctly allocated. However, the estimate of $\mathbf{A}$ may not be very reliable.

Fortunately, it can be proved that any convergent method of maximizing the likelihood function (9), in particular that part of it depending on $\mathbf{A}$, must yield an $\mathbf{A}^+$ giving complete separation if such an $\mathbf{A}^+$ exists. For suppose that the sample point $\mathbf{x}$ from H_s has $p_{t\mathbf{x}} > p_{s\mathbf{x}}$ for some value of $\mathbf{A}$. Now $p_{s\mathbf{x}} + p_{t\mathbf{x}} \leqslant 1$, so $p_{s\mathbf{x}} \leqslant \frac{1}{2}$. The part of the likelihood function that depends on $\mathbf{A}$ is a constant plus L', where

$$L' = \prod_{s=1}^{k} \prod_{\mathbf{x} \in H_s} p_{s\mathbf{x}}.$$

If separation is possible the maximum value of L' is 1 but if any sampie point does not satisfy (35), at least one term in L' is less than or equal to $\frac{1}{2}$, so $L' \leqslant 1$. A convergent procedure for $\mathbf{A}$ must give $\mathbf{A}^+$ with $L' > \frac{1}{2}$ at some stage and this $\mathbf{A}^+$ must give complete separation. This incidentally gives a conservative test for separation that can be built into the maximum likelihood procedure. Day & Kerridge (1967) gave these results for $k = 2$.

5. A simulation study of the properties of maximum likelihood estimation of logistic discriminators

Sample points were generated from two multivariate normal distributions with equal dispersion matrices to investigate the separate sample logistic discrimination method developed in §3. The values $n_1 = n_2 = 50$ and $p = 10$ were chosen. Since the functions $p_{s\mathbf{x}}$ $(s = 1, \ldots, k)$ are invariant under translations and rotations of the axes in the x-space, the common dispersion matrix was taken to be the identity and the means were given by $\boldsymbol{\mu}_1^T = (\theta, 0, \ldots, 0)$ and $\boldsymbol{\mu}_2^T = (-\theta, 0, \ldots, 0)$. Three values were taken for θ, 0·5, 1·0 and 1·5, and twenty sets of values were simulated with each of these. With the above parameter values and $\boldsymbol{\Pi} = \boldsymbol{\Pi}^*$, the true value of the logistic parameter, $\boldsymbol{\alpha}_1^T = (0, \boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T) = (0, 2\theta, 0, \ldots, 0)$. For each simulation, the estimated value, $\hat{\boldsymbol{\alpha}}_1$, was calculated using the method of §§3 and 4 with the true value as starting value. Convergence was obtained directly in all but two cases where restarting from the intermediate 'best' point, with the correct $\mathbf{F}^{-1}$ at that point, gave convergence. Complete separation only occurred with $\theta = 1·5$ and then in 6 out of 20 simulations.

Table 1. *The simulation of misclassification probabilities for logistic discriminators*

		Estimated discriminators								
		pr ($H_1 \to H_2$)				pr ($H_2 \to H_1$)				
		True		Observed		True		Observed		Theoretical best misclassification probabilities
θ	Number of simulations	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	
0·5	20	0·34	0·05	0·25	0·05	0·35	0·04	0·27	0·06	0·31
1·0	20	0·18	0·03	0·13	0·04	0·19	0·03	0·14	0·05	0·16
1·5	20	0·09	0·03	0·04	0·03	0·10	0·02	0·04	0·03	0·07
1·5 (no separation)	14	0·08	0·03	0·05	0·03	0·10	0·02	0·06	0·03	—
1·5 (separation)	6	0·12	0·03	0·00	0·00	0·98	0·02	0·00	0·00	—

Table 2. *Estimation of logistic parameters*

		α_{10}		α_{11}		α_{12}		α_{14}		α_{16}		α_{18}		$\alpha_{1,10}$	
θ	Number of simulations	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
0·5	20	−0·00 (0·00)	0·14	1·27 (1·00)	0·32	0·02 (0·00)	0·31	−0·28 (0·00)	0·27	0·67 (0·00)	0·30	−0·85 (0·00)	0·26	−0·41 (0·00)	0·33
1·0	20	0·08 (0·00)	0·39	2·71 (2·00)	0·30	−0·04 (0·00)	0·42	0·10 (0·00)	0·48	−0·05 (0·00)	0·43	−0·15 (0·00)	0·47	0·00 (0·00)	0·09
1·5	20	−0·85 (0·00)	3·79	35·50 (3·00)	35·11	−6·72 (0·00)	29·13	−17·14 (0·00)	24·95	9·67 (0·00)	41·27	1·97 (0·00)	3·67	−6·14 (0·00)	28·95
1·5 (no separation)	14	0·40	1·14	6·35	5·29	−0·34	1·07	0·70	1·36	−0·19	1·12	0·84	2·15	0·50	1·02
1·5 (separation)	6	−3·76	6·05	103·5	197·1	−21·6	53·31	−58·76	143·8	32·69	74·7	4·02	5·54	−21·65	52·6

The true values of the coefficients α_{1i} are shown in parentheses beneath the sample means.

The efficacy of the method for discrimination purposes can be judged by examining the probabilities of misclassification. Each estimate of the logistic discriminant function defines an allocation rule with true probabilities of misallocation, pr $(H_1 \rightarrow H_2)$ and pr $(H_2 \rightarrow H_1)$, and their observed values, that is the actual proportion of sample cases incorrectly assigned. The means and standard deviations of the 20 values of these parameters for each θ are noted in the first half of Table 1, together with the theoretical optimum misallocation probabilities, taking $\boldsymbol{\alpha}_1^T = (0, 2\theta, 0, \ldots, 0)$, as above.

It had been thought that the occurrence of sample points with complete separation would lead to poorer estimates of the discriminant function. This point is investigated in the second half of Table 1 where the results for $\theta = 1{\cdot}5$ are divided into the two groups: separation and no separation.

It can be seen in Table 1 that the method of estimating rules for discrimination introduced in § 3 works well in this context. The mean true misallocation probabilities are reasonably close to the optimum and the variability is not too large. As expected, the observed values of these probabilities are too optimistic to give anything but a rough guide. The effect of complete separation is less marked than anticipated, so that the allocation rule based on samples with this property is still useful.

Quite a different issue is whether good estimates of the coefficients $\boldsymbol{\alpha}_1$ are provided by the separate sample logistic discrimination method. Twenty estimated values of $\boldsymbol{\alpha}_1$ were available for each value of θ, so the sample means and dispersion matrices of these were calculated. The means and standard deviations are shown in the first half of Table 2. Since the sampling distributions of the estimates of $\alpha_{1j}\,(j = 2, \ldots, 10)$ are all the same, only a selection of these values is given. This shows that the estimated sampling distribution is reasonable for $\theta = 0{\cdot}5$ and $1{\cdot}0$, although perhaps a little biased in the latter case. However, for $\theta = 1{\cdot}5$, it is clear that the estimates are wild. In the second half of Table 2 these results are divided according to whether or not there was separation. As expected, the results with no separation are far better, although still not very good. Cox (1970) pointed out that with $k = 2$ and sampling from the mixture, the estimates of $\boldsymbol{\alpha}_1$ are not very good if

$$|\boldsymbol{\alpha}_1^T \mathbf{x}| > 3 \tag{38}$$

for a substantial proportion of the sampled points. The same is true of the situation here, and as θ goes from 0·5 to 1·5 the separation between the populations increases with the result that (38) is true for more and more of the sample points. Day & Kerridge (1967) also drew attention to this problem and suggested a two-tier sampling scheme from the mixture, the first being as before but the second being effectively from that part of the sample space where $|\boldsymbol{\alpha}_1^T \mathbf{x}| < 3$. Where feasible, this should be very effective and applicable to separate sampling.

The conclusion is that the method introduced in this paper is a good general method of discrimination but that the estimated values of the logistic coefficients are not reliable unless the condition (38) or its equivalent for $k > 2$ is untrue at a number of sample points.

6. The differential diagnosis of kerato-conjunctivitis sicca

There is a risk that people suffering from rheumatoid arthritis will also contract *kerato-conjunctivitis sicca*. This disease can be diagnosed reliably by an ophthalmic specialist but his services are not available to screen all rheumatoid arthritic patients. The question is whether a simple screening system can be devised to enable the medical staff of a rheumatic centre, who are not ophthalmic specialists, to decide which rheumatoid arthritic patients to refer to the eye hospital. It was thought that logistic discriminators might be useful in this context.

The diagnostic system was to be based on 10 symptoms of the presence or absence type. The ith observation, x_i, was taken to be 0 if the symptom was absent; otherwise it was taken to be 1. These observations were available on 40 rheumatoid arthritic patients with *kerato-conjunctivitis sicca* and 37 rheumatoid arthritic patients without *kerato-conjunctivitis sicca* 'normals'. This set of patients will be called Series I. The maximum likelihood estimate of α_1, $\hat{\alpha}_1$, was found iteratively using the approach of §4 with starting values given by Cox's (1970, p. 70) method. Let $z = (1, \mathbf{x}^T)\hat{\alpha}_1$, then

$$z = 4{\cdot}0 - 4{\cdot}4x_1 - 2{\cdot}1x_2 - 1{\cdot}1x_3 - 4{\cdot}7x_4 - 3{\cdot}5x_5 - 0{\cdot}8x_6 + 0{\cdot}8x_7 - 2{\cdot}4x_8 + 1{\cdot}8x_9 - 0{\cdot}9x_{10}. \quad (39)$$

The scores of all patients in Series I were calculated using equation (39) and are shown in Fig. 1 (*a*). It is clear that complete separation did not occur.

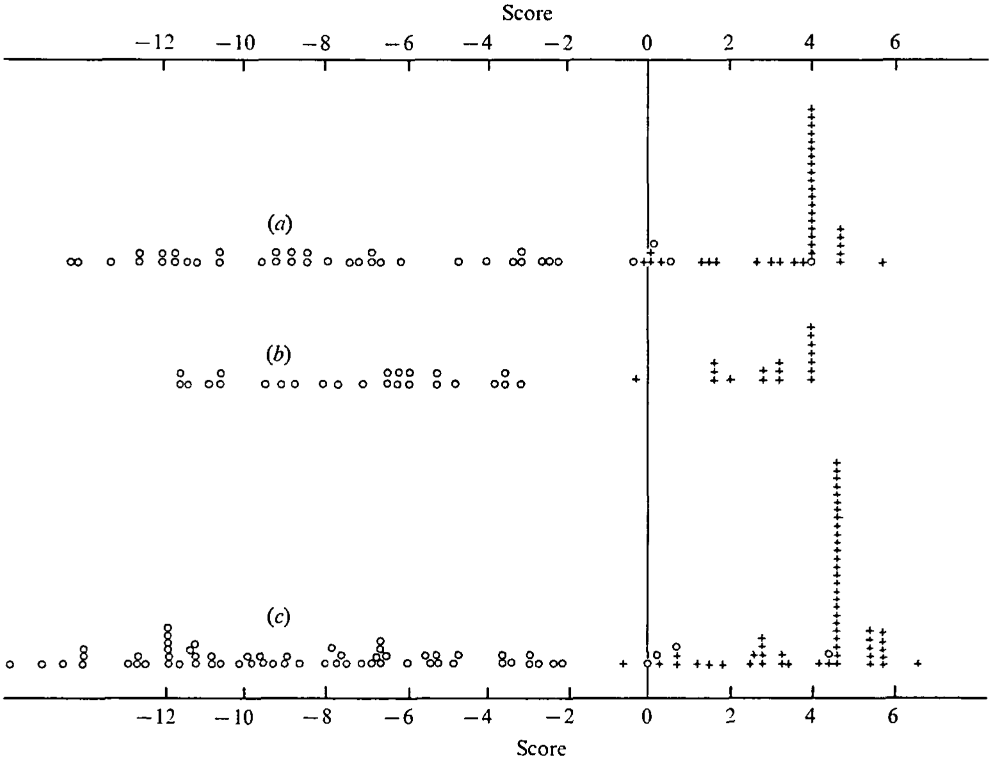

Fig. 1. The distribution of scores of *kerato-conjunctivitis sicca* patients (○) and normals (+): (*a*) Series I patients estimated from Series I; (*b*) Series II patients estimated from Series II; (*c*) Series I + II patients estimated from series I + II.

Doubts about diagnosis occur when a patient's score is small so it was decided to give a patient a queried diagnosis if his score was in the range −2 to 2. This corresponds approximately to odds between 9:1 in favour of *kerato-conjunctivitis sicca* and 9:1 against *kerato-conjunctivitis sicca*. This results in the following diagnostic system. Calculate z, then if

$$\begin{aligned} z &\geqslant 2\text{: call patient normal,} \\ -2 < z &< 2\text{: query diagnosis,} \\ -2 &\leqslant z\text{: diagnose } \textit{kerato-conjunctivitis sicca}. \end{aligned} \quad (40)$$

To test the diagnostic method estimated from the Series I patients, the ten symptoms were observed on a further set of 41 patients, Series II, which included 17 normals and 24 cases of *kerato-conjunctivitis sicca*. The scores of all these patients were calculated using equation (39) and are shown in Fig. 1 (*b*). It can be seen that the results in the two series are quite comparable.

For applications of this diagnostic method to further patients, the coefficients $\boldsymbol{\alpha}_1$ must be estimated from the largest possible sample. Thus, Series I and II were combined and the maximum likelihood estimate of $\boldsymbol{\alpha}_1$, $\hat{\boldsymbol{\alpha}}_1^{(2)}$, was found using the same method as before on the larger sample. Let $z_2 = (1, \mathbf{x}^T)\,\hat{\boldsymbol{\alpha}}_1^{(2)}$, then

$$z_2 = 4{\cdot}7 - 5{\cdot}2x_1 - 3{\cdot}0x_2 - 1{\cdot}3x_3 - 5{\cdot}4x_4 - 4{\cdot}0x_5 + 1{\cdot}1x_6 + 0{\cdot}8x_7 - 1{\cdot}9x_8 + 2{\cdot}1x_9 - 2{\cdot}0x_{10}. \tag{41}$$

The scores of all patients in Series I and II were calculated using equation (41) and these are shown in Fig. 1 (*c*). Again complete separation did not occur. A summary of all the results illustrated in Fig. 1, categorized as correct, query or wrong, is given in Table 3. It is concluded from Fig. 1 (*a*, *b*, *c*) and Table 3 that the results of diagnosing Series I and II patients using (41) are very similar. In addition, (39) and (41) are very much alike. Thus the diagnostic system given by (39), (40) and (41) is stable and repeatable. Moreover, the error and query rates are acceptable so the system is satisfactory from many points of view. Full details of this study are given by Anderson, Whaley, Williamson & Buchanan (1972).

Table 3. *Evaluation of the logistic discrimination method of the diagnosis of kerato-conjunctivitis sicca in rheumatoid arthritis*

Discriminator estimated from Series I

	Kerato-conjunctivitis sicca			No *kerato-conjunctivitis sicca*		
	Correct	Query	Wrong	Correct	Query	Wrong
Series I	36	3	1*	30	7	0
Series II	24	0	0	13	4	0

Discriminator estimated from Series I and II

	Kerato-conjunctivitis sicca			No *kerato-conjunctivitis sicca*		
	Correct	Query	Wrong	Correct	Query	Wrong
Series I+II	60	3	1*	47	7	0

* This patient had no symptoms.

7. Discussion

As mentioned in §2, the chief advantage of the logistic approach to discrimination is that the same technique can be used under many different assumptions about the underlying distributions. It is perhaps most useful when the observations are wholly or partly polychotomous. In the weakness of assumptions made about the $f_s(\mathbf{x})$, the method approaches the distribution-free techniques, for example, Fisher's linear discriminant function. However, unlike these, it also gives estimates of likelihood ratios and posterior probabilities. In some circumstances this could be the major objective of an investigation. If so, care must be taken if complete separation is thought possible as poor estimates of the $\boldsymbol{\alpha}_s$

and hence of the posterior probabilities or likelihood ratios would be likely. As mentioned in § 5, a two-stage sampling plan could mitigate these effects.

Given that the logistic discrimination approach is to be taken, sometimes both mixture and separate sampling are possible. The asymptotic properties of the two dispersion matrices do not help to make the choice since they are so similar (§ 3·2). However, it is the author's opinion that separate sampling should be chosen, provided that the n_s can be pre-selected to be approximately equal, because then balance between the n_s is guaranteed. It is conjectured that for a given total sample size n, samples with this balance give better estimates, on average, than those with imbalance. The latter is almost certain to occur with mixture sampling if one or more of the populations has low incidence. Unfortunately since the small-sample properties of maximum likelihood estimators are intractable, the above conjecture must remain as such.

To take a concrete example, suppose that retrospective sampling is planned from two diseases, D_1 and D_2, with relative incidences known to be 10 % and 90 %. Case histories are available, classified by diagnosis, and there is an annual total of about 100 cases. If resources only permit 100 histories to be examined, the mixture approach would take the last available year and find, say, $10 D_1$ and $90 D_2$ cases. However, the separate sampling plan would take $10 D_1$ and $10 D_2$ cases from each of the last 5 years, the D_2 cases selected randomly from the total in each year. This gives good balance between the populations and between the years.

Although the logistic model (5) has been related to simple allocation rules, using the posterior probability or likelihood ratio directly (§ 2), it can also be used with more complex techniques. For example, if it is required to maximize the expected utility, then $\mathbf{x}$ is allocated to action A_j (of $A_1, \ldots, A_v$) if

$$U_j \geqslant U_t \quad (t = 1, \ldots, v;\ t \neq s), \tag{42}$$

where

$$U_t = \sum_{s=1}^{k} \pi_s f_s(\mathbf{x})\, u_{st} \quad (t = 1, \ldots, v)$$

and u_{st} is the utility of A_t with H_s (Rao, 1965). It is clear that condition (42) can easily be written in terms of the $\{p_{s\mathbf{x}}\}$. Thus, given the $\{u_{st}\}$, the method of logistic discrimination, which yields estimates of the $\{p_{s\mathbf{x}}\}$, can be used to implement a decision-making system. Similarly the methods of § 3 can be used to furnish estimates of posterior probabilities required in constrained decision systems (Marshall & Olkin, 1968; Anderson, 1969). It is concluded that methods of estimating the logistic form of posterior probabilities (5) furnish a potentially valuable tool for a wide range of discrimination and decision problems.

The author is very grateful to Dr W. W. Buchanan of the Centre for Rheumatic Diseases, Glasgow, and Dr K. Whaley and Dr J. Williamson of the Western Infirmary, Glasgow, for providing the dataon rheumatic patients for the application of logistic discrimination in § 6.

References

Aitchison, J. & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29**, 813–28.

Anderson, J. A. (1969). Discrimination between k populations with constraints on the probabilities of misclassification. *J. R. Statist. Soc.* B **31**, 123–39.

Anderson, J. A., Whaley, K., Williamson, J. & Buchanan, W. W. (1972). A statistical aid to the diagnosis of *kerato-conjunctivitis sicca. Quarterly J. Med.* **41**. To appear.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. R. Statist. Soc.* B **25**, 220–33.

Cox, D. R. (1966). Some procedures associated with the logistic qualitative response curve. In *Research Papers in Statistics: Festschrift for J. Neyman*, ed. F. N. David, pp. 55–71. New York: Wiley.

Cox, D. R. (1970). *The Analysis of Binary Data.* London: Methuen.

Day, N. E. & Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics* **23**, 313–23.

Marshall, A. W. & Olkin, I. (1968). A general approach to some screening and classification problems. *J.R. Statist. Soc.* B **30**, 407–43.

Rao, C. R. (1965). *Linear Statistical Inference and its Applications.* New York: Wiley.

[*Received April* 1971. *Revised July* 1971]