A General Maximum Likelihood Discriminant

N. E. Day; D. F. Kerridge

Biometrics, Vol. 23, No. 2. (Jun., 1967), pp. 313-323.

A GENERAL MAXIMUM LIKELIHOOD DISCRIMINANT

N. E. DAY[1] AND D. F. KERRIDGE[2]

Research Group in Biometric Medicine, University of Aberdeen, Scotland

SUMMARY

A method of discrimination, based on maximum likelihood estimation, is described. On a variety of mathematical models, including and extending the models most commonly assumed in discriminant theory, the discriminant reduces to multivariate logistic analysis. Even when no simple model can be assumed, other considerations show that this method should work well in practice, and should be very robust with respect to departures from the theoretical assumptions. The method is compared with others in its application to a diagnostic problem.

1. INTRODUCTION

We discuss the problem of discriminating between only two populations, H_1 and H_2 , since the two-population problem is the one which arises most frequently in practice. Our results may be generalised without difficulty to the problem of more than two populations, should the need for this arise. We assume that an individual is drawn at random from either H_1 or H_2 , and a sample vector $\mathbf{X}$ composed of k separate variables x_i $(j = 1 \cdots k)$ is observed for this individual. The problem of discrimination is to produce an index based on $\mathbf{X}$ which helps to classify the individual as belonging to population H_1 or H_2 .

The most widely used method of discrimination, introduced by Fisher [1938], assumes that $\mathbf{X}$ is distributed according to the multivariate normal distribution, the covariance matrix being the same whether $\mathbf{X}$ arises from H_1 or H_2 , but the vector of means being different. This will be referred to as 'model F'. Another common method (Warner *et al.* [1961]), which has found considerable favour among statisticians engaged on diagnostic applications, assumes that each variable may take only two values, corresponding to a sympton recorded simply 'present' or 'absent', and that the separate variables are independent. This will be referred to as 'model W'. Although both methods of discrimination have been used successfully in many problems they are

[1] Present address: Department of Mathematical Statistics, Australian National University, Canberra, A.C.T., Australia.

[2] Present address: Department of Statistics, University of Aberdeen, Meston Walk, Old Aberdeen, Scotland.

not always suitable. It is unusual, at least in our experience, for all the components of the vector **X** to be approximately normally distributed, as model F requires. It would be even more surprising if the separate observations were independent, as model W requires.

There are of course many examples of sound statistical practices based on crude approximations, and our first reaction might well be to ignore these objections. However, there are special reasons to suppose that the effects of departures from the assumptions are not negligible in discrimination problems. The individuals we are most concerned with in this type of problem are those who are most difficult to classify. Generally these correspond to sample vectors which are not close to the mode of the distribution of **X** in either population. It is precisely at such extreme sample points that the departures from theoretical assumptions are most likely to have a serious effect.

The method of discrimination to be described in this paper is relevant to a very wide class of models, including both F and W and many generalisations of them. Even if no theoretical model is assumed, it is possible to show that, provided linear discrimination is adequate, the method should work reasonably well.

2. A GENERAL THEORETICAL MODEL

Let **A** be the variance-covariance matrix of a multivariate normal population, and let $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ be the mean vectors corresponding to populations H_1 and H_2 respectively. Finally let $\phi(\mathbf{X})$ be a non-negative scalar function of **X**. Then our general model for the density function of **X** is

$$dp = \alpha_i \exp \{-\tfrac{1}{2}(\mathbf{X} - \mathbf{\Lambda}_i)'\mathbf{A}^{-1}(\mathbf{X} - \mathbf{\Lambda}_i)\}\phi(\mathbf{X})\, d\mathbf{X} \qquad (i = 1, 2). \qquad (1)$$

We assume only sufficient restriction on the function $\phi(\mathbf{X})$ (e.g. integrability and non-negativeness) to ensure that this is a proper probability mass function; α_i is a normalising constant chosen to make the total probability unity.

If $\phi(\mathbf{X})$ is bounded, the model has a probabilistic interpretation as follows. We may re-normalise so that $\phi(\mathbf{X})$ lies in the range [0, 1]. Then we may regard **X** as a random variable generated from a multivariate normal population by von Neumann's [1961] rejection technique, $1 - \phi$ being the probability of rejection.

It will be clear that, on this general model, we may choose the distribution of **X** on H_1 say, with complete freedom. Having done this, however, we have only the choice of $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$, and **A** remaining to make the distribution of **X** on H_2 fit the form we require.

3. SOME IMPORTANT SPECIAL CASES

One obvious special case is obtained by taking $\phi(\mathbf{X}) \equiv 1$, when model F is recovered. A less obvious case is obtained as follows

(i) Let $\phi(\mathbf{X}) = 1$ if every component x_i of $\mathbf{X}$ is either 0 or 1
$= 0$ in all other cases

(ii) Let $\mathbf{A}$ be the identity matrix.

The second condition implies that each of the k components x_i of $\mathbf{X}$ is independent. For each component the relative probability of 0 or 1 is in the ratio of two ordinates of the normal probability density curve, at a distance of one standard deviation from each other. Clearly, since the mean may be freely chosen, this ratio may take any positive value. We have thus obtained model W as a second special case of the general model.

As an immediate generalisation of this model, we may drop the condition (ii) and allow $\mathbf{A}$ to be any possible variance-covariance matrix. It may be verified that although the variance-covariance matrix of $\mathbf{X}$ will differ from $\mathbf{A}$, it is completely arbitrary. This special case of the general model has important applications to diagnostic problems, since it enables us to represent systems of correlated symptoms.

Another useful special case may be obtained by applying the condition (i) to only a restricted subset of the components x_i. This yields a model of the situation in which some variables are continuous, while others are dichotomous.

One application to skew distributions may also be of interest. In medical applications, a variable is often approximately normally distributed in 'control' patients, but definitely skew in the 'abnormal' group. Commonly the tail of the distribution in the 'abnormals' furthest from the mean of the 'controls' is exaggerated compared to the normal model. This situation may be met by making $\phi(\mathbf{X})$ constant in the region of the 'control' mean, but letting it increase towards and beyond the mean of the 'abnormal' group.

The examples of models obtained by special choices of the function $\phi(\mathbf{X})$ should be sufficient to illustrate the breadth of application of the general model. The great advantage of the present approach is that we do not need, in practice, to decide on any particular form of $\phi(\mathbf{X})$, since the form of the function has no influence on the solution. The presence of this undetermined function in the model makes for great robustness in the method.

4. THE OPTIMAL DISCRIMINANT

If individuals selected at random from some super-population belonged to H_1 or H_2 with known relative frequency, and the distribution of $\mathbf{X}$ on H_1 and H_2 were exactly known, the form of the optimal discriminant would be determined in an obvious fashion by the application of Bayes's Theorem (Hoel and Petersen [1949]). This is a strictly frequentist and universally acceptable application of Bayes's Theorem. We have

$$P(H_i \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid H_i)P(H_i)}{\sum_{i=1,2} P(\mathbf{X} \mid H_i)P(H_i)}$$

and it is clear that no procedure can make fewer errors of classification than that of choosing H_1 or H_2 according to which is more frequently true on the observed $\mathbf{X}$. In general, however, neither the frequencies $P(H_i)$ nor the exact form of $P(\mathbf{X} \mid H_i)$ are exactly known. Under these circumstances it seems reasonable to attempt to estimate $P(H_i \mid \mathbf{X})$ on the basis of an observed sample.

The assumption that definite prior frequencies $P(H_i)$ exist, even though they may be unknown, calls for further discussion, as it may be unrealistic in some cases. Even if definite frequencies do not exist, we might argue on the basis of decision theory that it is reasonable to behave as if they did. However, this point will not be urged too strongly, as in many problems, including those of greatest interest to us, the existence of such frequencies is quite clear. For example, H_1 might be the population of individuals with a particular disease, and H_2 the population free from it. In such a case $P(H_i)$ is the incidence of the disease, which is meaningful for any sufficiently clearly defined population.

The argument may require slight modification if the relative frequency of H_1 is not the same in the population from which the observed sample is drawn as it is in the population to which the discriminant is to be applied. The adjustment to be made to the discriminant in such a case will be discussed later.

Applying the general model described previously, we find

$$P(H_1 \mid \mathbf{X}) = \frac{P(H_1)\alpha_1 \exp\{-\tfrac{1}{2}(\mathbf{X} - \mathbf{\Lambda}_1)'\mathbf{\Lambda}^{-1}(\mathbf{X} - \mathbf{\Lambda}_1)\}\phi(\mathbf{X})}{\sum_{i=1,2} P(H_i)\alpha_i \exp\{-\tfrac{1}{2}(\mathbf{X} - \mathbf{\Lambda}_i)'\mathbf{\Lambda}^{-1}(\mathbf{X} - \mathbf{\Lambda}_i)\}\phi(\mathbf{X})}$$

which simplifies to give

$$P(H_1 \mid \mathbf{X}) = \frac{\exp(\mathbf{X}'\mathrm{B} + c)}{1 + \exp(\mathbf{X}'\mathbf{B} + c)} \tag{2}$$

where

$$\mathbf{B} = \mathbf{\Lambda}^{-1}(\mathbf{\Lambda}_1 - \mathbf{\Lambda}_2)$$

and

$$c = -\tfrac{1}{2}(\mathbf{\Lambda}_1 - \mathbf{\Lambda}_2)\mathbf{\Lambda}^{-1}(\mathbf{\Lambda}_1 + \mathbf{\Lambda}_2) + \log\left\{\frac{P(H_1)\alpha_1}{P(H_2)\alpha_2}\right\}. \tag{3}$$

The linear function $\mathbf{X}'\mathbf{B} + c$ is positive when H_1 is more probable, and so corresponds to the more familiar form of the linear discriminant. For some purposes, however, it is convenient to be able to give the result in the form of an estimated conditional probability. For example, in medical diagnostic applications, it is useful to be able to classify patients attending a clinic into the three or more categories such as 'abnormal', 'normal' and 'doubtful'. The simple expression of the posterior probability in terms of a logistic function may be regarded as a generalisation of a result due to Cornfield *et al.* [1961] for the univariate normal population.

5. THE APPLICATION OF MAXIMUM LIKELIHOOD ESTIMATION

It is clear that if $\phi(\mathbf{X})$ were known exactly, we could apply maximum likelihood estimation in a straightforward fashion to estimate $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}$, and combine these estimates to give estimates of $\mathbf{B}$ and c. For example, when $\phi \equiv 1$ (model F) we obtain

$$\hat{\mathbf{B}} = \mathbf{V}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$$

$$\hat{c} = -\tfrac{1}{2}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)\mathbf{V}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) + \ln (n_1/n_2)$$

where $\bar{\mathbf{X}}_1$, $\bar{\mathbf{X}}_2$, $\mathbf{V}$ represent the sample values of the mean vectors and (pooled) covariance matrix, and n_1 , n_2 are the numbers from the populations H_1 and H_2 in the sample. Apart from the term involving n_1/n_2, this is a standard form of Fisher's discriminant function. We may note that this provides a way of calculating $P(H_i \mid \mathbf{X})$ from the ordinary Fisherian discriminant.

When $\phi(\mathbf{X})$ is unknown, which is the case of immediate interest, we are faced with the difficulty of maximizing a likelihood function containing an unknown function.

The complete likelihood function is

$$\begin{aligned} L' &= \prod_{r \epsilon S_1} P(\mathbf{X}_r \mid H_1)P(H_1) \times \prod_{r \epsilon S_2} P(\mathbf{X}_r \mid H_2)P(H_2) \\ &= \prod_{r \epsilon S_1} \alpha_1 P(H_1)\phi(\mathbf{X}_r) \exp\left[-\tfrac{1}{2}(\mathbf{X}_r - \mathbf{A}_1)'\mathbf{A}^{-1}(\mathbf{X}_r - \mathbf{A}_1)\right] \\ &\quad \times \prod_{r \epsilon S_2} \alpha_2 P(H_2)\phi(\mathbf{X}_r) \exp\left[-\tfrac{1}{2}(\mathbf{X}_r - \mathbf{A}_2)'\mathbf{A}^{-1}(\mathbf{X}_r - \mathbf{A}_2)\right], \end{aligned}$$

where S_1 , S_2 denote the set of values of r for which the sample vector $\mathbf{X}_r$ arises from the populations H_1 , H_2 respectively. Then using the result (2) of section 4

$$\begin{aligned} L' &= \prod_{r \epsilon S_1 + S_2} P(\mathbf{X}_r) \times \prod_{r \epsilon S_1} P(H_1 \mid \mathbf{X}_r) \times \prod_{r \epsilon S_2} P(H_2 \mid \mathbf{X}_r) \\ &= \prod_{r \epsilon S_1 + S_2} \phi(\mathbf{X}_r) \sum_{i = 1,2} \alpha_i P(H_i) \exp\left[-\tfrac{1}{2}(\mathbf{X}_r - \mathbf{A}_i)'\mathbf{A}^{-1}(\mathbf{X}_r - \mathbf{A}_i)\right] \\ &\quad \times \prod_{r \epsilon S_1} \frac{\exp(\mathbf{X}_r'\mathbf{B} + c)}{1 + \exp(\mathbf{X}_r'\mathbf{B} + c)} \times \prod_{r \epsilon S_2} \frac{1}{1 + \exp(\mathbf{X}_r'\mathbf{B} + c)}. \end{aligned}$$

Since $P(H_1)$, $P(H_2)$ and $\phi(\mathbf{X})$ may be chosen freely, we may make the first factor achieve its maximum (which occurs when the $P(\mathbf{X}_r)$ are equal for all r) for any values of $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}$. The maximum does not depend on these quantities, while the rest of the likelihood can be maximized by varying *only* $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}$. We therefore ignore the first product and estimate $\mathbf{B}$ and c by maximizing

$$L = \prod_{r \epsilon S_1} \frac{\exp(\mathbf{X}_r'\mathbf{B} + c)}{1 + \exp(\mathbf{X}'\mathbf{B} + c)} \times \prod_{r \epsilon S_2} \frac{1}{1 + \exp(\mathbf{X}'\mathbf{B} + c)}.$$

This argument rests on the assumption that $\phi(\mathbf{X})$ can be chosen with complete freedom at every point $\mathbf{X}_r$ in the observed sample. Even if some information about the form of $\phi(\mathbf{X})$ is available, the procedure remains valid, (though not necessarily fully efficient) since we may observe that L is in fact a likelihood function, composed of the conditional frequencies $P(H_i \mid \mathbf{X}_r)$ and as such maximum likelihood may properly be applied to it. Our examination of the unconditional likelihood function suggests that the loss of information, if any, which this conditional inference involves, is likely to be small unless the information about ϕ is very precise. Considerable doubts have been expressed about the assumption, for example, of multivariate normality on discriminant analysis (see for example, the discussion following Hills' [1966] paper). Hence leaving ϕ as an arbitrary function disposes, at least in part, of a serious objection to the usual theoretical analysis. We then have

$$\log L = n_1(\bar{\mathbf{X}}_1'\mathbf{B} + c) - \sum_{r \epsilon S_1 + S_2} \log [1 + \exp(\mathbf{X}_r'\mathbf{B} + c)].$$

From this we may determine $\mathbf{B}$ and c by an iterative procedure. This tends to be lengthy if the number of variables is large, but presents no difficulty on a computer. (An autocode programme for an Elliot 803 computer is available on request.)

In some cases, the discriminant is to be applied to the classification of a future population in which the frequencies of H_1, H_2 are known to be $P^*(H_1)$, $P^*(H_2)$, respectively, different from the frequencies $P(H_1)$, $P(H_2)$ in the population from which the sample is drawn. Since $\mathbf{B}$ and c for the theoretically optimal discriminant are given by formulae (3), $\hat{\mathbf{B}}$ is unaffected, and the estimate $\hat{c}$ obtained from the sample need only be changed to

$$\hat{c} + \log (n_2 P^*(H_1))/(n_1 P^*(H_2)).$$

6. A LOSS FUNCTION APPROACH

The justification so far put forward for the form of discriminant suggested rests on a particular, though general, theoretical model. To some extent all theoretical models are suspect from a practical point of view, and an alternative approach may be interesting. We may write

$$-\log L = \sum_{r \varepsilon S_1} \log (1 + e^{-z_r}) + \sum_{r \varepsilon S_2} \log (1 + e^{z_r})$$

where $z_r = \mathbf{X}_r'\mathbf{B} + c$. We should therefore have a discrimination rule identical with that derived from the maximum likelihood theory if we considered that the 'cost' of assigning a score z_r to an individual was $\log (1 + e^{-z_r})$ or $\log (1 + e^{z_r})$, according as the individual arose from H_1 or H_2, and attempted to minimise the cost on this basis. The cost of assigning a large negative score to one individual from H_2 (leading to a correct classification) is thus near zero. As z_r increases the cost increases monotonically, being asymptotically equal to z_r for large positive scores (corresponding to a mis-classification). It will be seen that only individuals who are near the borderline or are actually misclassified make any appreciable contribution to the total cost, and hence have much influence on the estimates $\hat{\mathbf{B}}$ and $\hat{c}$. This does seem to correspond to what we want in many practical problems. Much previous theoretical work on discrimination has treated the discriminant as a decision maker. If this is really what we want, as it certainly will be in many industrial problems, for example, the cost function which has been described is not a natural one. If the discriminant index is to be the sole basis for an irreversible decision, the cost may depend only on whether a mistake is made or not. If, on the other hand, the value of the discriminant index is used as a general guide, possibly with other information, the cost should depend on the extent to which the index is misleading. For example, in applications to medical diagnosis, there is always the possibility of making further tests if the index falls into a doubtful zone, and of supplementing the index with clinical judgement. Under such circumstances the present approach is clearly relevant.

In this cost function approach we make no assumption about the distributions involved, except for the very mild assumption that a discriminant which minimises the cost function in an observed sample is likely to have low cost when applied to a future sample. We may therefore regard the method of this paper as non-parametric. A cost function analysis roughly along these lines was in fact developed by us independently of the maximum likelihood approach, until it was found that the two ideas could be combined.

7. COMPLETE SEPARATION OF POPULATIONS

In some cases it may be possible to find a discriminant function which classifies the $\mathbf{X}_r$ in the observed sample with complete success into those arising from H_1, and those arising from H_2. Although of course the success of a discriminant in classifying the sample on which it is based does not by any means guarantee an equal success in a future sample, it would seem desirable to find such a discriminant if it exists. It may be proved that if a linear discriminant with this property can be constructed, the method of this paper will always find it. No other method known to us at present can be guaranteed to do this.

Let $z_r^* = \mathbf{X}_r'\mathbf{B}^* + c^*$ be a discriminant which completely separates the two populations in the observed sample. Without loss of generality we may assume that every value of z_r^* arising from H_1 is positive, and every value arising from H_2 is negative. Let $z_r = \mathbf{X}_r'\mathbf{B} + c$ be any discriminant function which does not separate the populations completely. Then either $z_r < 0$ for some r in S_1 or $z_r > 0$ for some r in S_2, since otherwise complete separation would have been achieved.

We may write the likelihood function in the form

$$L = \prod_{r \epsilon S_1} \left(\frac{e^{z_r}}{1 + e^{z_r}} \right) \times \prod_{r \epsilon S_2} \left(\frac{e^{-z_r}}{1 + e^{-z_r}} \right).$$

The logistic function

$$y = e^z/(1 + e^z)$$

increases monotonically from $y = 0$ at $z = -\infty$ to $y = 1$ at $z = 0$ and approaches 1 as $z \rightarrow +\infty$. In the case of the second discriminant, therefore, L is the product of positive terms, each of which is less than or equal to unity, and at least one is $\frac{1}{2}$ or less. Hence $L \leq \frac{1}{2}$. Consider now the likelihood function, for arbitrary θ

$$L^*(\theta) = \prod_{r \epsilon S_1} \left(\frac{e^{\theta z_r^*}}{1 + e^{\theta z_r^*}} \right) \times \prod_{r \epsilon S_2} \left(\frac{e^{-\theta z_r^*}}{1 + e^{-\theta z_r^*}} \right).$$

Since every term in both products tends to one as θ increases, $L^*(\theta)$ will certainly exceed $\frac{1}{2}$ for some θ. Hence we have found a new discriminant function for which the corresponding likelihood is greater than is possible for any discriminant which does not separate the populations completely. It follows that any convergent process for maximizing the likelihood must eventually find a discriminant which separates the population if one exists. That the actual estimates of $\mathbf{B}$ and c would tend to infinity if the process were not artificially terminated leads to no difficulty in practice, since it is easy to recognize that complete separation has been achieved if L is calculated at each stage. This is advisable in any case, so that appropriate action may be taken if the iterative process leads to a decrease in L at any stage.

8. UPDATING A DISCRIMINANT

A discriminant function needs modifying from time to time, even after it has been put into regular service. Two reasons are important: the conditions may have changed slightly, so making the original discriminant out of date, or the original discriminant may have been based on relatively small numbers, so that more precise estimation of the optimum discriminant is desirable. With most methods of discrimination, the only way to do this seems to be to take a new random sample and classify each member of it exactly, as was done in setting up the discriminant in the first place. This is both time-consuming and expensive. A great practical advantage of this method of discrimination is that information obtained in the course of the everyday use of the discriminant can be used to update it.

We may take for example the case of a discriminant used as a screening procedure in clinical diagnosis. Patients with a score above a certain level will be regarded as almost certainly having a particular disease: those with scores below a certain level as almost certainly not having it. The 'doubtfuls' with intermediate scores are referred for further investigation and a definite diagnosis is finally reached. We then carry out our discrimination procedure on these fully investigated patients alone. It is clear that the restriction on the sample space affects the parameters $\phi(\mathbf{X})$, α_1, α_2 and $P(H_1)$ and $P(H_2)$ of the general model, but not $\mathbf{A}_1$, $\mathbf{A}_2$ or $\mathbf{A}$. Denote by an asterisk the new parameters of the model for the problem in the restricted sample space, which we denote by D, and let $\theta_i = P(\mathbf{X}_i \varepsilon D \mid H_i)$. Then

$$\begin{aligned} \phi^*(\mathbf{X}) &= \phi(\mathbf{X}) \quad \text{if} \quad \mathbf{X} \varepsilon D \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Since the total probability integral of $\mathbf{X}$ must be unity, we have

$$\alpha_1^*/\alpha_2^* = \alpha_1(1 - \theta_2)/\alpha_2(1 - \theta_1),$$

and

$$P^*(H_1)/P^*(H_2) = P(H_1)(1 - \theta_1)/P(H_2)(1 - \theta_2)$$

is the ratio of patients from the two populations among the patients fully investigated.

Hence for the optimal discriminant

$$\mathbf{B}^* = \mathbf{A}^{-1}(\mathbf{A}_1 - \mathbf{A}_2) = \mathbf{B}$$

$$c^* = -\tfrac{1}{2}(\mathbf{A}_1 - \mathbf{A}_2)\mathbf{A}^{-1}(\mathbf{A}_1 + \mathbf{A}_2) + \log\left\{\frac{P^*(H_1)\alpha_1^*}{P^*(H_2)\alpha_2^*}\right\} = c.$$

It is therefore perfectly valid to use the information about the doubtfuls alone to calculate a discriminant function for use on the whole sample space. We may also note that even if all the patients had been referred for complete investigation, the 'doubtful' region would include nearly all the points which have much influence on the estimates $\hat{\mathbf{B}}$ and $\hat{c}$. Hence we have incidentally achieved a form of stratified sample with high efficiency. This in itself may be of great practical value in other applications.

There is no reason, in theory at least, why we should not update the discriminant after every new observation, by adding it to the original sample, although for practical reasons a periodic updating is likely to be preferable.

9. QUADRATIC DISCRIMINATION

Reverting to the general model, it may be seen that if we allowed the covariance matrices in the two populations to be unequal, the argument of the logistic function would be a quadratic instead of a linear form. Most of the preceding results generalise to cover this case, but the estimation problem would be rather unwieldy unless the number of variables were very small.

10. AN EXAMPLE

This method of discrimination was compared with the methods of discrimination based on models F and W (see section 1), in application to a diagnostic problem. The aim was to construct a linear discriminant, based on signs and symptoms recorded at two hypothyroid clinics. It was hoped that using such a discriminant, many patients could be classified as normal or abnormal without the need for laboratory tests.

The sample consisted of the entire intake of patients over many months at the two clinics: only those who had previously been treated for thyroid disease were excluded. All were submitted to laboratory tests and it was found that 56 patients out of the total intake of 152 were suffering from thyroid disease. Discriminants of each type were constructed on the basis of the same ten signs and symptoms.

A reasonable way to use such a discriminant, produced by any of the methods, is to classify any future patient with a higher score than that observed in a normal individual in the sample as 'probably abnormal'. Similarly a patient with a lower score than that observed in any abnormal individual is classified 'probably normal'. Those with intermediate scores are classified 'doubtful' and should be referred for laboratory investigation.

Applying this rule to the original sample, we found that using the method of discrimination discussed in this paper, there were 85 patients in the 'doubtful' category. Using the methods based on models F and W there were 93 and 105 'doubtfuls' respectively. The superiority of the logistic method in this particular problem is not unexpected on theoretical grounds, since the variables do not satisfy either of the models F or W. It is true that the proportion of 'doubtfuls' in the original sample will tend to underestimate the proportion in a future sample. However, as the sample is large, and the number of variables small, there is no reason to suppose that this bias is large, and it seems unlikely that it could account for much of the observed superiority of the logistic discriminant over the others.

REFERENCES

Cornfield, J., Gordon, T. and Smith, W. [1961]. Quantal Response Curves for Experimentally Uncontrolled Variates. *Bull. Inter. Stat. Inst. 38*, 97–115.

Fisher, R. A. [1938]. The Statistical Utilisation of Multiple Measurements. *Ann. Eug., Lond., 7*, 179–88.

Hills, M. [1966]. Allocation rules and their error rates. *J. R. Statist Soc. B28*, 1–31.

Hoel, P. and Petersen [1949]. A solution to the problem of optimum classification. *Ann. Math. Statist. 20*, 433–38.

von Neumann, J. [1961]. Various Techniques used in Connection with Random Digits. Paper No. 13 in *Monte Carlo Methods*. National Bureau of Standards, Applied Mathematics Series No. 12. Washington: U. S. Govt. Printing Office.

Warner, H., Toronto A., Veesey L. and Stephenson R. [1961]. A Mathematical Approach to Medical Diagnosis. *J. Amer. Med. Asn. 177*, 177–83.