A SOLUTION TO THE PROBLEM OF OPTIMUM CLASSIFICATION

By P. G. Hoel and R. P. Peterson

University of California, Los Angeles

1. Summary. By means of a general theorem, the space of the variables of classification is separated into population regions such that the probability of a correct classification is maximized. The theorem holds for any number of populations and variables but requires a knowledge of population parameters and probabilities. A second theorem yields a large sample criterion for determining an optimum set of estimates for the unknown parameters. The two theorems combine to yield a large sample solution to the problem of how best to discriminate between two or more populations.

2. Introduction. There are essentially two basic problems in discriminant analysis. The first problem is to test whether the populations differ, since it would be futile to attempt a classification if the populations did not differ. The second problem is to find an efficient method for classifying individuals into their proper populations. In this paper, an optimum asymptotic solution of the second problem will be presented.

3. Parameters known. Let $f_i = f_i(x_1, \cdots, x_k)$, $(i = 1, \cdots, r)$ denote the probability density function of population i in the region under consideration. Let $p_i > 0$, $(i = 1, \cdots, r)$, denote the probability that population i will be sampled if a single individual is selected at random from that region, and let R denote the k dimensional Euclidean variable space. Then the desired theorem is the following:

Theorem 1. *If M_i denotes the region in R where $p_i f_i \geq p_j f_j$, $(j = 1, \cdots, r)$, and where $p_i f_i > 0$, then the set of regions M_i, $(i = 1, \cdots, r)$, in which any overlap is assigned to the M_i with the smallest index, will maximize the probability of a correct classification.*

For the purpose of proving this theorem, consider any other set of non-overlapping regions, M_i'. Since the addition to any of the regions M_i of a part of R throughout which all the functions f_i vanish will not affect the probability of a correct classification, there is no loss of generality in assuming that the set of regions M_i' contains the same portion of R as the set of regions M_i does. The relationship between the two sets may be expressed by means of the formulas

$$M_i = \sum_{j=1}^{r} M_{ij} \tag{1}$$

and

$$M_j' = \sum_{i=1}^{r} M_{ij}, \tag{2}$$

where M_{ij} denotes that part of M_i which is contained in M_j'.

Since a sample point that falls in the region M_i will be judged to have come from population i, the probability of the correct classification of a single random sample by means of the set M_i is given by

$$Q = p_1 \int_{M_1} f_1 dE + \cdots + p_r \int_{M_r} f_r dE, \tag{3}$$

where $dE = dx_1 dx_2 \cdots dx_r$. If Q' denotes the probability of the correct classification by means of the set M_i',

$$Q' = p_1 \int_{M_1'} f_1 \, dE + \cdots + p_r \int_{M_r'} f_r \, dE.$$

In the notation of (1) and (2), these probabilities become

$$Q = p_1 \int_{\sum_j M_{1j}} f_1 \, dE + \cdots + p_r \int_{\sum_j M_{rj}} f_r \, dE$$

and

$$Q' = p_1 \int_{\sum_i M_{i1}} f_1 \, dE + \cdots + p_r \int_{\sum_i M_{ir}} f_r \, dE.$$

Now consider the difference $Q - Q'$. It can be expressed in the form

$$\begin{aligned} Q - Q' &= \sum_{i=1}^{r} \sum_{j=1}^{r} \left[p_i \int_{M_{ij}} f_i \, dE - p_j \int_{M_{ij}} f_j \, dE \right] \\ &= \sum_{i=1}^{r} \sum_{j=1}^{r} \int_{M_{ij}} [p_i f_i - p_j f_j] \, dE. \end{aligned}$$

Since M_{ij} is contained in M_i and $p_i f_i \geq p_j f_j$, $(j = 1, \cdots, r)$, holds throughout M_i, it follows that each of these integrals is non-negative; consequently $Q \geq Q'$, which proves the theorem.

This theorem yields a solution to the classification problem only when the f_i are completely specified and the p_i are known.

It will be observed that this theorem is similar to a generalization of a fundamental lemma in the Neyman-Pearson theory of testing hypotheses [1], and to a result by Welch [2].

If the basic weight function in Wald's [3] formulation of the multiple decision problem assumes only the values 0 and 1, corresponding to whether or not a correct classification is made, it will be found that the set of regions M_i will minimize the expected value of the loss in that formulation.

4. Parameters unknown. Since the p_i, as well as the parameters in the f_i, are assumed to be unknown, Q will be a function of such parameters. Let $\theta_1, \cdots, \theta_s$ denote all such parameters, including the p_i. Now let a random sample of size n be taken from the region under consideration and let $\bar{\theta}_1, \cdots, \bar{\theta}_s$ denote a set of estimates of the parameters based on this sample. Since the total sample will constitute a sample of size n_1 from f_1, n_2 from f_2, etc., where $n = n_1 + \cdots + n_r$, the θ's for f_i will be estimated by means of a sample of size n_i rather than of size n. In the following arguments, it will not be necessary to distinguish between θ's which are estimated by different size samples because the arguments will be based on the order of terms with respect to the size sample and $n_i \sim np_i$ with probability one. Or, more simply, choose all n_i equal.

Let $\bar{M}_i$ correspond to M_i when the parameters are replaced by their sample estimates and let $\bar{Q}$ denote the probability of a correct classification when using the regions $\bar{M}_i$ in place of the regions M_i. Then, from (3),

$$Q - \bar{Q} = \sum_{i=1}^{r} p_i \left[\int_{M_i} f_i \, dE - \int_{\bar{M}_i} f_i \, dE \right].$$

Let $H = Q - \bar{Q}$. Since the estimates, $\bar{\theta}_i$, are random variables, H will be a random variable which is a function of the estimation functions, $\bar{\theta}_i$, as well as of the parameters, θ_i. The desired criterion for determining optimum estimates is then given by the following theorem:

THEOREM 2. *If $E(\bar{\theta}_i - \theta_i)^4 = O(n^{-g})$, $g > 0$, and if in some neighborhood of the point $\bar{\theta}_i = \theta_i$, $(i = 1, \cdots, s)$ the function H is continuous and possesses continuous derivatives of the first, second, and third order with respect to the $\bar{\theta}_i$, then*

$$E(H) = \frac{1}{2} \sum_{i=1}^{s} \sum_{j=1}^{s} H_{ij} E(\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j) + O(n^{-3/4g}),$$

where H_{ij} denotes the partial derivative of H with respect to $\bar{\theta}_j$ and $\bar{\theta}_i$ at the point $(\theta_1, \cdots, \theta_s)$.

The proof is similar to the type of proof used by Cramer [4] to obtain an expression for the variance of a function of central moments.

By means of Tchebycheff's inequality [4], page 182, it follows that

$$P[(\bar{\theta}_i - \theta_i)^4 \geq \epsilon^4] \leq \frac{E(\bar{\theta}_i - \theta_i)^4}{\epsilon^4}.$$

From the theorem assumptions, there exists a constant A such that

$$P[\bar{\theta}_i - \theta_i)^4 \geq \epsilon^4] < \frac{An^{-g}}{\epsilon^4}.$$

This is equivalent to

$$P[|\bar{\theta}_i - \theta_i| \geq \epsilon] < \frac{An^{-g}}{\epsilon^4}.$$

If E_1 denotes the set of points in sample space where $|\bar{\theta}_i - \theta_i| < \epsilon$, $(i = 1, \cdots, s)$, and E_2 denotes the complementary set, this inequality implies that

$$P[E_2] < \frac{sAn^{-g}}{\epsilon^4}. \tag{4}$$

The expected value of H may be written in the form

$$E(H) = \int_{E_1} H\,dP + \int_{E_2} H\,dP. \tag{5}$$

Consider the order of the second integral. From (4) and the fact that H is the difference of two probabilities, it follows that

$$\left| \int_{E_2} H\,dP \right| \leq \int_{E_2} dP = P[E_2] < \frac{sAn^{-g}}{\epsilon^4}.$$

Consequently (5) becomes

$$E(H) = \int_{E_1} H\,dP + O(n^{-g}). \tag{6}$$

Now consider the first integral. From the theorem assumptions, if ϵ is chosen sufficiently small, it follows that for any point in the set E_1, the function H can be expanded in the form

$$H = H(\theta) + \sum_1^s (\bar{\theta}_i - \theta_i)H_i(\theta) + \frac{1}{2}\sum_1^s \sum_1^s (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)H_{ij}(\theta) + R,$$

where θ denotes the point $(\theta_1, \cdots, \theta_s)$, where

$$R = \frac{1}{6}\sum_1^s \sum_1^s \sum_1^s (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)(\bar{\theta}_k - \theta_k)H_{ijk}(\theta'),$$

and where θ' is some point in E_1. Since $\bar{Q}$ reduces to Q when $\bar{\theta} = \theta$, $H(\theta) = 0$. Furthermore, since Q denotes the maximum probability of a correct classification, $H \geq 0$ for all $\bar{\theta}$; hence $H_i(\theta) = 0$ and $H_{ii}(\theta) \geq 0$ for all i. Thus, for any point in the set E_1,

$$H = \frac{1}{2}\sum_1^s \sum_1^s (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)H_{ij}(\theta) + R.$$

If this expression is substituted in (6), $E(H)$ will become

$$E(H) = \frac{1}{2}\sum_1^s \sum_1^s H_{ij}(\theta) \int_{E_1} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)\,dP + \int_{E_1} R\,dP + O(n^{-g}). \tag{7}$$

Consider, first, the order of the remainder term. From the continuity assumption on H_{ijk}, it follows that H_{ijk} is bounded in E_1, say $| H_{ijk}(\theta') | < B$; hence

$$\left| \int_{E_1} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)(\bar{\theta}_k - \theta_k)H_{ijk}(\theta')\,dP \right| < B \int_{E_1} | (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)(\bar{\theta}_k - \theta_k) |\,dP.$$

By Schwarz's inequality,

$$\int_{E_1} | (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)(\bar{\theta}_k - \theta_k) |\,dP$$
$$\leq \left[\int_{E_1} (\bar{\theta}_i - \theta_i)^2(\bar{\theta}_j - \theta_j)^2\,dP \int_{E_1} (\bar{\theta}_k - \theta_k)^2\,dP \right]^{\frac{1}{2}}.$$

Similarly,

$$\int_{E_1} (\bar{\theta}_i - \theta_i)^2(\bar{\theta}_j - \theta_j)^2\, dP \leq \left[\int_{E_1} (\bar{\theta}_i - \theta_i)^4\, dP \int_{E_1} (\bar{\theta}_j - \theta_j)^4\, dP\right]^{\frac{1}{2}},$$

$$\int_{E_1} (\bar{\theta}_k - \theta_k)^2\, dP \leq \left[\int_{E_1} (\bar{\theta}_k - \theta_k)^4\, dP \int_{E_1} dP\right]^{\frac{1}{2}} \leq \left[\int_{E_1} (\bar{\theta}_k - \theta_k)^4\, dP\right]^{\frac{1}{2}}.$$

Since

$$\int_{E_1} (\bar{\theta}_i - \theta_i)^4\, dP \leq \int_{E_1+E_2} (\bar{\theta}_i - \theta_i)^4\, dP = O(n^{-g}),$$

the preceding inequalities combine to give

$$\left|\int_{E_1} R\, dP\right| = O(n^{-3/4g}). \tag{8}$$

Now consider the first integral in (7). It may be written in the form

$$\int_{E_1} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)\, dP = E(\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j) - \int_{E_2} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)\, dP. \tag{9}$$

By Schwarz's inequality,

$$\left|\int_{E_2} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)\, dP\right| \leq \left[\int_{E_2} (\bar{\theta}_i - \theta_i)^2\, dP \int_{E_2} (\bar{\theta}_j - \theta_j)^2\, dP\right]^{\frac{1}{2}}.$$

Similarly,

$$\int_{E_2} (\bar{\theta}_i - \theta_i)^2\, dP \leq \left[\int_{E_2} (\bar{\theta}_i - \theta_i)^4\, dP \cdot P[E_2]\right]^{\frac{1}{2}}.$$

If these inequalities are combined and inequality (4) is employed, (9) will reduce to

$$\int_{E_1} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)\, dP = E(\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j) + O(n^{-g}). \tag{10}$$

Finally, if (8) and (10) are employed in (7), it will reduce to the result stated in the theorem.

The order of the leading term in $E(H)$ depends upon the nature of the estimating functions, $\bar{\theta}_i$. In order to insure that this term will be the dominating term, and thus rule out pathological situations, only that class of estimating functions (estimators) will be considered for which this term will be of lower order than that of the remainder term. If the estimators are means or central moments, for example, then $g = 2$. For such estimators the order of the remainder term is $O(n^{-\frac{3}{2}})$, whereas the order of the leading term is not higher than $O(n^{-1})$.

A set of estimators will be called an optimum set if it maximizes the expected value of the probability of a correct classification, or, what is equivalent, if it minimizes $E(H)$. Since only large samples are being considered here, it is necessary to define optimum in an asymptotic sense. Consider sets of estimators for which $E(H)$ is of order $O(n^{-q})$. For this class of estimators, a set will be called asymptotically optimum if it minimizes

$$\lim_{n\to\infty} n^q E(H).$$

Among asymptotically optimum sets of various orders, the set corresponding to the highest order would naturally be considered as the best asymptotic set. Now from Theorem 2, it readily follows that a set of estimators which minimizes

$$\sum_1^s \sum_1^s H_{ij} E(\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j) \tag{11}$$

will be an asymptotically optimum set.

5. Maximum likelihood estimates. If the estimates $\bar{\theta}_i$ are unbiased and uncorrelated, (11) will reduce to

$$\sum_1^s H_{ii}\sigma_i^2 \tag{12}$$

where $\sigma_i^2 = E(\bar{\theta}_i - \theta_i)^2$ is a function of n as well as of the parameters. Since, from the discussion preceding (7), $H_{ii} \geq 0$, it follows that (12) will be a minimum when the σ_i^2 assume their minimum values. Now it is known [4], page 504, that under mild restrictions maximum likelihood estimates possess minimum asymptotic variances; hence for estimators of the type being considered which also satisfy the conditions in [4], the maximum likelihood estimates of the θ_i will yield an asymptotically optimum set of estimates for the classification problem.

REFERENCES

[1] J. NEYMAN AND E. S. PEARSON, "On the problem of the most efficient tests of statistical hypotheses," *Roy. Soc. Phil. Trans.*, Vol. 231 (1933), pp. 289–337.

[2] B. L. WELCH, "Note on discriminant functions," *Biometrika*, Vol. 31 (1939), pp. 218–220.

[3] A. WALD, "Contributions to the theory of statistical estimation and testing hypotheses," *Annals of Math. Stat.*, Vol. 10 (1939), pp. 299–304.

[4] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, 1946, pp. 352–356.