THEORIES AND METHODS IN CLASSIFICATION: A REVIEW*

by

Somesh Das Gupta**

Technical Report No. 201

March 1973

University of Minnesota
Minneapolis, Minnesota

*A summary of this review was presented at the Symposium on Discrimination and Classification at Athens, Greece, in 1972.

**Supported by U.S. Army Research Grant DA-ARO-D-31-124-70-G-102.

Theories and Methods in Classification: A Review*

Somesh Das Gupta**
University of Minnesota

*A summary of this review was presented at the Symposium on Discrimination and Classification at Athens, Greece, in 1972.

**Supported by U.S. Army Research Grant DA-ARO-D-31-124-70-G-102.

1A. Introduction.

In this review paper I have restricted my attention only to major theoretical papers. However I have tried to be objective, as far as possible, in selecting papers from the enormous bulk of literature in this area. The recently published bibliography on multivariate analysis (Anderson, Das Gupta, and Styan, 1972) lists over 400 papers published before 1967 in the area of classification and discrimination. Moreover, some results are available in the well-known textbooks by Anderson (1958) and by Rao (1952), besides few books (listed in the references) completely devoted to this and allied fields. Anyway, I apologize for omitting many papers, especially many important applied papers and useful computer programs.

In the literature we find many names for this general area of problems; for example, allocation, identification, prediction, pattern recognition, selection, besides the standard terms, such as, classification and discrimination. Whatever names may be attached, it is clear that this branch has attracted many researchers from different disciplines. From the existing literature, I have extracted the main formulations of the classification problem and reviewed almost all the important results under different broad categories of problems.

1B. Early History

In the first survey of discriminatory analysis, Hodges (1950) aptly mentioned the following.

> In his invited address at the meeting of the Institute of Mathematical Statistics in Berkeley, California, June 16, 1949, Professor M. A. Girshick pointed out that the development of discriminatory analysis reflects the same broad phases as does the general history of statistical inference. We may distinguish a Pearsonian stage,..., followed by a Fisherian stage. Professor Girshick further notes a Neyman-Pearson stage and a contemporary Waldian stage....

In the early work, the classification problem was not precisely formulated and often confounded with the problem of testing the equality of two or more distributions; the term "discriminatory analysis" was used for both. In practice, the following scheme was generally followed for the two-population classification problem. Suppose we have three distributions F, F_1, F_2 and T_i is a test statistic designed to test the hypothesis $F = F_i$ $(i = 1, 2)$. The decision $F = F_i$ is taken if T_i is the smaller of T_1 and T_2; sometimes the critical values of T_i's are compared in order to take the decision. Thus statistics for testing the equality of two distributions played an important role. Generally, such a test statistic may be considered as a measure of divergence between the two distributions. Karl Pearson (in a paper by Tildesley (1921)) proposed one such measure, termed as the "coefficient of racial likeness (CRL)." This was modified by Morant in 1928 and by Mahalanobis in 1927 and 1930. Mahalanobis called his measure D^2 and suggested (1930) also some measures of divergence in variability, skewness and kurtosis and studied the distributions of these measures. In 1926, Pearson published the first considerable theoretical work on the CRL and suggested the following form for the coefficient when the variables are dependent:

$$\frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2),$$

where $\bar{x}_i$ is the sample mean vector based on a sample of size n_i from the i^{th} population $(i = 1, 2)$ and S is the pooled sample covariance matrix. In 1935 and 1936, Mahalanobis gave the dependent-variate versions of his D^2-statistic in the classical and the studentized forms.

The distributions of these statistics were studied by Bose (1936 a, b), Bose (1936), Bose and Roy (1938), and Bhattacharya and Naryana (1941). In 1931 Hotelling suggested a test statistic T^2 which is a constant multiple of the studentized Mahalanobis D^2 and obtained its null distribution.

Hodges remarded that "the first clear statement of the problem of discrimination, and the first proposed solution to that problem were given by Fisher in the middle of the 1930's...the ideas of Fisher first appeared in print in papers by other people (Barnard (1935), Martin (1936))." Earlier than this Morant (1926) considered the problem of classifying a skull into Eskimo or modern English groups by two sets of tests. Fisher's own first work on the subject appeared in his paper in 1936. For the univariate two-population classification problem Fisher suggested a rule which classifies the observation x into the i^{th} population if $|x-\bar{x}_i|$ is the smaller of $|x-\bar{x}_1|$ and $|x-\bar{x}_2|$. For p-component observation vector ($p > 1$), Fisher reduced the problem to the univariate one by considering an "optimum" linear combination (called the "linear discriminant function") of the p components. For a given linear combination Y of the p components, Fisher considered the ratio between the difference in the sample means of the Y-values and the standard error within samples of the Y-values and maximized this ratio in order to define the optimum linear combination. It turns out that the coefficients of this optimum linear combination are proportional to $S^{-1}(\bar{x}_1-\bar{x}_2)$. Incidentally, Fisher (1936) suggested a test for the equality of two normal distributions with the same unknown covariance matrix and this test is the same as the one proposed by Hotelling (1931).

The next development was influenced by the pioneering fundamental work by Neyman and Pearson (1933, 1936). For the two-population problem, Welch (1939)

derived the forms of Bayes rules and the minimax Bayes rule when the distributions are known; he illustrated the theory with multivariate normal distributions with the same covariance matrix. This example was also considered by Wald (1944) who further proposed some heuristic rules by replacing the unknown parameters by their respective (maximum likelihood) estimates. Wald studied the distribution of the proposed classification statistic. Von Mises (1944) obtained the rule which maximizes the minimum probability of correct classification. The problem of classification into two normal distributions with different covariance matrices was treated by Cavalli (1945) and Penrose (1947) when $p = 1$ and by Smith (1947) for general p. In a series of papers, Rao (1946, 1947a, 1947b, 1948, 1949a, 1949b, 1950) suggested different methods of classification into two or more populations following the ideas of Neyman-Pearson and Wald; in particular, Rao suggested a measure of distance between two groups and considered the possibility of withholding decision (through "doubtful" regions) and preferential decisions. Rao's development is for the case when the distributions are all known. General theoretical results on the classification problem (as a special case) in the framework of decision theory are given in the book by Wald (1950) and in a paper by Wald and Wolfowitz (1950).

References (1)

Tildesley, M. L. (1921). A first study of the Burmese skull. Biometrika 13 247-251.

Pearson, Karl (1926). On the coefficient of racial likeness. Biometrika 18 105-117.

Morant, G. M. (1926). A first study of crainology of England and Scotland from neolithic to early historic times, with special reference to Anglo-Saxon skulls in London museums. Biometrika 18 56- .

Mahalanobis, P. C. (1927). Analysis of race mixture in Bengal. Jour. and Proc. Asiatic Soc. of Bengal. 23 No. 301-333.

Morant, G. M. (1928). A preliminary classification of European races based on cranical measurements. Biometrika 20(B) 301-375.

Mahalanobis, P. C. (1930). On tests and measurements of group divergence. Jour. and Proc. Asiatic Soc. Bengal. 26 541-588.

Hotelling, H. (1931). The generalization of Student's ratio. Ann. Math. Statist. 2 360-378.

Neyman, J. and Pearson, E. S. (1933a). On the problem of the most efficient tests of statistical hypotheses. Phil. Trans. Roy. Soc. A 231 281-

Neyman, J. and Pearson, E. S. (1933b). On the testing of statistical hypotheses in relation to probability a priori. Proc. Camb. Phil. Soc. 9 492-

Barnard, M. M. (1935). The secular variations of skull characters in four series of Egyptian skulls. Ann. Eug. 6 352-371.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proc. Nat. Inst. Sci. India 2 49-55.

Bose, R. C. (1936a). On the exact distribution and moment coefficients of the D^2-statistic. Sankhyā 2 143-154.

Bose, R. C. (1936b). A note on the distribution of differences in mean values of two samples drawn from two multivariate normally distributed populations, and the definition of the D^2-statistic. Sankhyā 2 379-384.

Neyman, J. and Pearson, E. S. (1936). Contributions to the theory of testing statistical hypotheses. I - Stat. Res. Memoir, London, 1, 1-37.

Bose, S. N. (1936). On the complete moment coefficients of the D^2-statistics. Sankhyā 2 385-396.

Martin, E. S. (1936). A study of the Egyptian series of mandibles with special reference to mathematical methods of sexing. Biometrika 28 149-178.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eug. 7 179-188.

Fisher, R. A. (1938). The statistical utilization of multiple measurements. Ann. Eugen. 8 376-386.

Bose, R. C. and Roy, S. N. (1938). The distribution of the studentized D^2-statistics. Sankhya 4 19-38.

Welch, B. L. (1939). Note on discriminant functions. Biometrika 31 218-220.

Goodwin, C. N. and Morant, G. M. (1940). The human remains of Iron Age and other periods from Maiden Castle, Dorset. Biometrika 31 295-

Bhattacharya, D. P. and Narayan, R. D. (1941). Moments of the D^2-statistic for populations with unequal dispersions. Sankhyā 5 401-412.

Day, B. B. and Sandomire (1942). Use of the discriminant function for more than two groups. J. Amer. Statist. Assoc. 37 461-472.

Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. Ann. Math. Statist. 15 145-162.

von Mises, R. (1945). On the classification of observation data into distinct groups. Ann. Math. Statist. 16 68-73.

Cavalli. L. L. (1945). Alumni problemi della analisi biometrica di popolazioni naturali. Mem. Ist. Idrobiol. 2 301-323.

Rao, C. R. (1946). Test with discriminant functions in multivariate analysis. Sankhya 7 407-413.

Rao, C. R. (1947a). The problem of classification and distance between two populations. Nature 159 30-31.

Rao, C. R. (1947b). A statistical criterion to determine the group to which an individual belongs. Nature 160 835-836.

Brown, G. W. (1947). Discriminant functions. Ann. Math. Statist. 18 514-528.

Penrose, L. S. (1947). Some notes on discrimination. Ann. Eug. 13 228-237.

Smith, C. A. B. (1947). Some examples of discrimination. Ann. Eug. 13 272-282.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. Jour. Roy. Stat. Soc. (B) 10 159-203.

Aoyama, H. (1950). A note on the classification of data. Ann. Inst. Statist. Math. 2 17-20.

Rao, C. R. (1949a). On the distance between two populations. Sankhyā 9 246-248.

Rao, C. R. (1949b). On some problems arising out of discrimination with multiple characters. Sankhyā 9 343-366.

Wald, A. (1950). Statistical Decision Functions. Wiley, New York.

Rao, C. R. (1950). Statistical inference applied to classificatory problems. Sankhyā 10 229-256.

Hodges, J. L. (1950). Survey of discriminatory analysis. USAF School of Aviation Medicine. Rep. No. 1. Randolph Field, Texas.

Wald, A. and Wolfowitz, J. (1950). Characterization of the minimal complete class of decision functions when the number of decisions and distributions is finite. Proc. Second Berkeley Symp. on Stat. and Probability.

Books

Rao, C. R. (1952). Advanced Statistical Methods in Biometric Research (Chapters 8 and 9). Wiley, New York.

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis (Chap. 6). Wiley, New York.

Sebestyn, G. S. (1962). Decision making processes in pattern recognition MacMillan Co., New York.

Uhr, L. M. (1966). Pattern recognition theory: theory, simulations, and dynamic models of form perception and discovery. Wiley, New York.

Fu, K. S. (1968). Sequential Methods in Pattern Recognition and Machine Learning. Academic Press, New York.

Helstrom, C. W. (1968). Statistical Theory for Signal Detection. Pergamon Press, New York.

Watanabe, M. S. (ed.) (1972). Frontiers of Pattern Recognition. Academic Press, New York.

Patrick, E. A. (1972). Fundamentals of pattern recognition. Prentice-Hall, Englewood Cliffs, New Jersey.

Anderson, T. W., Das Gupta, S., and Styan, G. P. H. (1972). A Bibliography of Multivariate Statistical Analysis. Oliver & Boyd, Edinburgh.

2. Problems in classification; different situations.

The following are the main formulations of the classification problems along with some important variations.

Problem 1. Let $\pi_1, \pi_2, \ldots, \pi_m$ be m distinct populations (of experimental units). A random sample (of size $n \geq 1$, or a sequence) of experimental units is available from a population π_0 which is known to be exactly one of $\pi_1, \ldots, \pi_m$; the problem is to decide which one. In order to distinguish the populations a vector (of p components, $p \geq 1$, or a sequence of components) X of real-valued characteristics of a unit is considered. The distribution of X in π_i is denoted by P_i (the corresponding c.d.f. of the p-dimensional distribution being F_i), $i = 0, i, \ldots, m$. Since $P_0 = P_I$ for some $I = 1,2,\ldots,m$, the problem is to choose one among the m decisions $I = i (i=1,\ldots,m)$. A decision rule is based on the observations on the units to be classified and the available information on $P_1, \ldots, P_m$. If these distributions are not completely known, supplementary information on them is obtained through available samples from the populations $\pi_1, \ldots, \pi_m$. In this case, it is assumed that $F_1 \times F_2 \times \ldots \times F_m$ belongs to a certain set Ω of distributions and for each point $F_1 \times F_2 \times \ldots \times F_m$, F_i's are taken to be different. The samples from the populations $\pi_1, \ldots, \pi_m$ will be called "training" samples, (this term is used generally by engineers), and P_i's (or F_i's) will be called "class-distributions."

Problem 2. Here π_0 is considered to be a mixture of the populations $\pi_1, \ldots, \pi_m$. Corresponding to each unit we define X as before and consider a number I which denotes the serial number $(1,2,\ldots,m)$ of the population to which the unit belongs. For the units to be classified I is unobservable and the problem is to decide on the value of I from the

knowledge of X. The distribution of X, given $I = i$, is P_i and the distribution of I is over the set $\{1,2,\ldots,m\}$. The problem will be termed as "known mixture" or "unknown mixture" according as the distribution of I is known or unknown. If the distribution of (X,I) is unknown, a (training) sample from π_0 (of size $N \geq 1$, or a sequence) is available to get information on it. A training sample may be of two types (i) "Supervised" or "identified" --For each unit in the training sample both X and I are observable. (ii) "Unsupervised" or "unidentified"-- For each unit in the training sample only X is observable.

Sometimes the units to be classified occur in a sequence and after the ith unit is classified its exact I-value becomes available. Thus for classifying the nth unit, the previous $n - 1$ units form a supervised training sample. We shall call this case as (iii) post-supervised or post-identified.

It may also happen that the units to be classified do not come from the sample population, but, in a given sample, the number of units from each population may be known.

Problem 2G. In the above problem I is taken as a classificatory variable and it is artificial in nature. More generally, one may consider I as a continuous or discrete variable with physical meaning and the population π_i corresponds to $I \in S_i$, where $S_1,\ldots,S_m$ is a partition of the I-space. Marshall and Olkin (1968) incorporated the decision of observing I along with the m decisions in their formulation.

Instead of considering only the m decisions, one may also incorporate the possibility of reserving judgments, preferential decision, as well as consider a more general decision space as $P_0 \in \{P_{i_1},\ldots,P_{i_k}\}$ where $(i_1,\ldots,i_k)$

in a subset of $(1,2,\ldots,m)$ and $k = 1,2,\ldots, m-1$.

The populations $\pi_1, \pi_2,\ldots,\pi_m$ may represent m different "states" or points of time. In that case, one may get a training sample such that on each of its units X-observations are available at these m points of time. This would lead to "dependent" training sample.

Also it may not be possible to observe every component of X on each sampled unit. This would give rise to "incomplete" data. It may be mentioned that one may consider a general stochastic process instead of a finite-dimensional vector X.

The possibility of treating m as unknown is not considered in this review.

3. Classification Into Known Distributions: General Theories.

Suppose that the distributions of X are $P_1,\ldots, P_m$ in $\pi_1,\ldots, \pi_m$, respectively, and the p.d.f. of X in π_i is given by f_i with respect to a σ-finite measure μ. We shall first consider the problem of classifying one unit from π_0 into one of $\pi_1,\ldots, \pi_m$, from the decision-theoretic viewpoint. The problem can be stated as a zero-sum two-person game where each person has m possible actions. Let $\ell(i,j) \geq 0$ be the loss for classifying a unit into π_i when it really belongs to π_j; assume $\ell(i,i) = 0$ for all $i = 1,\ldots, m$. A decision rule is given by $\delta = (\delta_1,\ldots, \delta_m)$ where $\delta_i(x)$ is the conditional probability of classifying into π_i given the observation x. The risk-vector of a rule δ is given by $r(\delta) = (r_1(\delta),\ldots, r_m(\delta))$, where $r_j(\delta) = \sum_{i=1}^{m} \ell_{ij}\int \delta_i(x)dP_j(x)$. When $\ell_{ij} = \ell_j$ for all $i \neq j$, $r_j(\delta) = \ell_j\int[1-\delta_j(x)]dP_j(x) = \ell_j[1-\alpha_j(\delta)]$, where $\alpha_j(\delta)$ is the probability of correct classification (PCC) for the rule δ when π_j is the correct population. Correspondingly the probabilities of misclassification (PMC) are given by $\int \delta_i(x)dP_j(x)$, $i \neq j$.

A prior distribution is given by $\xi = (\xi_1,\ldots, \xi_m)$, where ξ_j is the probability that π_j is the true population. (In case of mixed population, the distribution of X can be expressed as $\sum_j \xi_j P_j$.) The ξ-Bayes risk of a rule δ is given by $R(\xi, \delta) = \sum_{j=1}^{m} \xi_j r_j(\delta)$. The main results are as follows.

(i) A necessary and sufficient condition for a rule δ to be ξ-Bayes is that for any j $(j = 1,\ldots, m)$, $\delta_j(x) = 0$ for all x (except possibly for a set of μ-measure 0) for which $L_j(x) > \min_{1\leq i\leq m} L_i(x)$, where $L_i(x) = \sum_{k=1}^{m} \ell_{ik}\xi_k f_k(x)$. In particular, when $\ell_{ij} = 1$ for all $i \neq j$, the above inequality can be expressed as $\xi_j f_j(x) < \max_{1\leq i\leq m} \xi_i f_i(x)$.

(ii) The class of all admissible rules is complete (and hence minimal complete).

(iii) Every admissible rule is Bayes.

(iv) For every prior distribution ξ, there exists an admissible ξ-Bayes rule.

(v) There exists a least favorable distribution ξ^0 and a minimax rule δ^M which is admissible and ξ^0-Bayes. For every minimax rule δ,

$$r_i(\delta) \leq R(\xi^0, \delta) = \max_{1 \leq i \leq m} r_i(\delta), \quad i = 1, \ldots, m.$$

(va) If $m = 2$, there exists a unique minimax rule (and hence admissible Bayes) for which $r_1(\delta^M) = r_2(\delta^M)$.

(vb) Suppose $\ell_{ij} = \ell > 0$ for $i \neq j$, and the distributions $P_1, \ldots, P_m$ are mutually absolutely continuous. Then there exists a unique minimax rule δ^M for which

$$r_1(\delta^M) = \ldots = r_m(\delta^M).$$

It can be shown that if either of the above two conditions is violated, the equality of the risk components of δ^M may not hold.

For proofs of the above results one may see Wald (1950, Section 5.1.1) and Ferguson (1967), although there are many papers and books [including Rao (1952), Anderson (1958)] which deal with this problem and present results weaker than the above. For earlier work see Welch (1939) and Von Mises (1945). Raiffa (1961) considered comparisons among experiments along standard lines dealing with risk functions.

A Bayes rule may lead to large PMC's and there have been several attempts to overcome this difficulty. Anderson (1969) posed the classification problem with $m + 1$ actions, the additional action being termed as a "deferred judgement" or "query." In that case, a decision rule δ is given by $(\delta_0, \delta_1, \ldots, \delta_m)$, where $\delta_0(x)$ is the conditional probability of suspending judgement. He considered the problem of maximizing $\sum_{i=1}^{m} \xi_i \alpha_i(\delta)$ subject to constraints given by

$$\int \delta_i(x) dP_j(x) \leq c_{ij}, \quad (i, j = 1, \ldots, m; \ i \neq j)$$

where c_{ij}'s are given constants, and obtained results on the existence,

necessity, sufficiency and uniqueness of such solutions. Neyman and Pearson (1936) dealt with the maximization of the PCC to one population subject to the PMC's of m other populations being equal to specified levels. See also Lehmann (1959). Rao (1952) also considered the problem posed by Anderson but gave only sufficient conditions. There is a heuristic discussion in Rao (1952) on introducing doubtful decisions (or regions) or preferential decisions besides the m actions. Quesenberry and Genaman (1968) considered the classification problem as a (2^m-1)-decision problem as follows:

$\delta_{i_1,\ldots,i_s}$: means decide that $P_0 \in \{P_{i_1},\ldots, P_{i_s}\}$ for $s = 1,\ldots, m-1$

δ_0: means reserve judgement,

where $(i_1,\ldots, i_s)$ is a subset of $(1,\ldots, m)$. They posed the problem of finding a rule which minimizes the probabilities of reserving judgement when the probabilities of wrong decisions (i.e., P_j (decide $P_0 \neq P_j$)) are controlled; they gave the solution for $m = 2$.

Marshall and Olkin (1968) (see Section 2 for their problem) gave some characterizations of minimum risk procedures. The possibility of observing the components of X sequentially was also considered. See Cochran (1951) for a related problem.

Heuristic Rules: A likelihood-ratio (LR) rule is defined by δ, where $\delta_j(x) > 0$, if (for some positive constants $c_1,\ldots, c_m$) $c_j f_j(x) < \max_{1\le i\le m} [c_i f_i(x)]$; see the result (a). In particular, if c_i's are all equal, the rule will be called a maximum-likelihood (ML) rule.

For a distance function d defined for pairs of distributions, a minimum distance (MD) rule classifies X into F_i if $d(\hat{F}_0, F_i) = \min_{1\le j\le m} d(\hat{F}_0, F_j)$; ties may be resolved in some manner. In the above, $\hat{F}_0$ is an estimate of F_0 obtained from the sample (from π_0) to be classified, so that $d(\hat{F}_0, F_j)$ are defined; when $F_0 = F(\cdot; \theta)$, one may consider $F_0 = F(\cdot; \hat{\theta})$.

Suppose we restrict our attention to rules belonging to a given class. Then one may find an optimum rule in that class (if it exists) by maximizing a weighted average of the PCC's or minimaxing PMC's. See Aoyama (1950) in this connection.

For $m = 2$, the classification problem is essentially the problem of testing a simple hypothesis against a simple alternative. In this case, there are well-known results on the asymptotic behavior of the error probabilities. See Kullback (1958), Chernoff (1952). For $m > 2$, see Hellman and Raviv (1970) Suppose $\epsilon(F_1, F_2, t)$ is the average error probability for a rule $X \lesseqgtr t$, when $m = 2$. Chernoff (1970) showed that

$$\sup_{F_i \epsilon \mathfrak{F}_i} \inf_{t} \epsilon(F_1, F_2, t) = [2(1 + \Delta^2)]^{-1},$$

where $\mathfrak{F}_i$ is the class of all univariate distributions with means μ_i and variance σ_i^2, and $\Delta = |\mu_1 - \mu_2|/(\sigma_1 + \sigma_2)$. The same result is obtained if one considers only the LR tests. For other studies on error probabilities, see Bahadur (1971) and references therein.

The problem of distinguishing (i.e., finding sequential or non-sequential rules so that the PMC's can be controlled arbitrarily) between two sets of distributions is posed by Hoeffding and Wolfowitz (1958) and some necessary and sufficient conditions were obtained by them. In this framework, one considers a sequence of i.i.d. random variables whose common distribution is known to belong to either of two given sets. Freedman (1967) extended some of these results when the possible distributions are countably many. Yarborough (1971) studied this problem with likelihood-ratio rules. The papers dealing with discrimination between stochastic processes are mainly concerned with finding conditions for which two (or more) processes (i.e., the induced measures) are equivalent or orthogonal. In case of equivalence, the next problem

is to obtain the likelihood-ratio and study some rules based on it. For Gaussian processes, see Feldman (1958), Hajek (1958), Rao and Varadarajan (1963) and the references therein. Brown (1971) dealt with Poisson processes; Shepp (1965), and Kantor's (1967) results are concerned with distinguishing between a process and its translate. For general work in this area, see Kakutani (1948), Gikhman and Skorokhod (1966), Kraft (1955), and Adhikari (1957). When we have a sample $(X_1,\ldots, X_{n_0})$ from π_0 and the problem is to make decisions on their common distribution (which is known to be one of $F_1,\ldots, F_m$) one may treat the problem from the standard decision-theoretic view-point. It is also possible to use the compound-decision approach of Robbins (1951) in order to get some asymptotically good rules. The empirical Bayes approach may be used when the observations are from a mixed population.

When one has the possibility of getting observations from π_0 sequentially, it may be appropriate to consider the sequential m-decision problem. Out of a considerable literature, the following may be mentioned: Wald (1947), Wald (1950), Armitage (1950), Mallows (1953), Simons (1967), Fu (1968, and the references therein), Meilijson (1969), Roberts and Mullis (1970), Kinderman (1972).

Dvoretzky, Kiefer and Wolfowitz (1953) pointed out that most of the results of Wald (1947) extend to the case of stochastic processes in continuous time provided that the last observation is a sufficient statistic for the entire past and the log(LR) of these statistics at various points of time form a process with stationary and independent increments; e.g., sequentially testing the drift of a Brownian motion or the intensity of a Poisson process. Bhattacharya and Smith (1972) defined sequential probability ratio tests for testing a simple hypothesis against a simple alternative for the mean value function of a real Gaussian process with known covariance kernel; exact formulas are obtained for the error probabilities and the OC function.

References (3).

Neyman, J. and Pearson, E. S. (1936). See Ref. 1.

Welch, B. L. (1939). See Ref. 1.

Von Mises, R. (1945). See Ref. 1.

Wald, A. (1947). Sequential Analysis. Wiley, New York.

Kakutani, S. (1948). On equivalence of infinite product measures. Ann. Math. 49 214-224.

Wald, A. (1950). Statistical Decision Functions. Wiley, New York.

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses and its relation to discriminant function analysis. J. Roy. Statist. Soc. Ser.B. 12 137-144.

Aoyama (1950). A note on the classification of observation data. Ann. Inst. Statist. Math. 2 17-19. (MR-12)

Cochran, W. G. (1951). Improvement by means of selection. Proc. II Berkeley Symp. Math. Statist. Prob. Univ. of California Press, 449-470.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on sum of observations. Ann. Math. Statist. 23 493-507.

Rao, C. R. (1952). See Ref. 1 (books).

Mallows, C. L. (1953). Sequential discrimination. Sankhyā 12 321-338.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1953). Sequential decision problems for processes with continuous time parameter testing hypotheses. Ann. Math. Statist. 24 254-264.

Kraft, C. (1955). Some conditions for consistency and uniform consistency of statistical procedures. Univ. Calif. Publ. Statist. 2 125-142.

Adhikari, B. P. (1957). Analyse discriminante des measures de probabilite sur un espace abstrait. C.R. Acad. Sci. 244 845-846 (MR-18).

Anderson, T. W. (1958). See Ref. 1 (books).

Feldman, J. (1958). Equivalence and perpendicularity of Gaussian processes. Pacific Jour. Math. 8 699-708.

Hajek, J. (1968). On a property of the normal distribution of any stochastic process. Czech. Math. J. 8 610-618.

Hoeffding, W. and Wolfowitz, J. (1958). Distinguishability of sets of distributions. Ann. Math. Statist. 29 700-718.

Kullback, S. (1959). Information Theory and Statistics. Wiley, New York (Dover-1968)

Lehmann, E. L. (1959). Testing Statistical Hypotheses. Wiley, New York.

Raiffa, H. (1961). Statistical Decision Theory approach to item selection for dichotomous tests and criterion variables. Stad. Item. Anal. Pred., Stanford University Press, 187-220.

Rao, C. R. and Varadarajan, V. S. (1963). Discrimination of Gaussian processes. Sankhyā Ser. A. 25 303-330.

Shepp, L. A. (1965). Distinguishing a sequence of random variables from a translate of itself. Ann. Math. Statist. 36 1107-1112.

Phatarford, R. M. (1965). Sequential analysis of dependent observations. Biometrika 52 157-165.

Gikhman, I. I. and Skorokhod, A. V. (1966). On the densities of probability measures in function spaces. Russian Math. Surveys 21 83-156.

Simons, G. (1967). Lower bounds for average sample number of sequential multihypothesis tests. Ann. Math. Statist. 38 1343-1365.

Ferguson, T. S. (1967). Mathematical Statistics: A Decision-Theoretic Approach. Academic Press, New York.

Freedman, D. A. (1967). A remark on sequential discrimination. Ann. Math. Statist. 38 1666-1676.

Marshall, A. W. and Olkin, I. (1968). A general approach to some screening and classification problems. J. Roy. Statist. Soc. Ser. B. 30 407-435.

Quesenberry, C. P. and Gessaman, M. P. (1968). See Ref. 7.

Fu, K. S. (1968). See Ref. 1 (books).

Anderson, J. A. (1969). Discrimination between k populations with constraints on the probabilities of misclassification. J. Roy. Statist. Soc. Ser. B. 31 123-139.

Kantor, M. (1969). On distinguishing translates of measures. Ann. Math. Statist. 40 1773-1777.

Meilijson, I. (1969). A note on sequential multiple decision procedures. Ann. Math. Statist. 40 653-657.

Chernoff, H. (1970). A bound on the classification error for discriminating between populations with specified means and variances. Stanford Univ. Dept. of Statistics Tech. Report No. 66.

Roberts, R. A. and Mullis, C. T. (1970). A Bayes sequential test of m hypotheses. IEEE Trans. on Information Science. IT-16 91-94.

Hellman, M. and Raviv, J. (1970). Probability of error, equivocation and the Chernoff bound. IEEE Trans. Inform. Theory IT-16 368-372.

Bahadur, R. R. (1971). Some Limit Theorems in Statistics. SIAM (Bristol, England).

Brown, M. (1971). Discrimination of Poisson processes. Ann. Math. Statist. 42 773-776.

Yarborough, D. A. (1971). Sequential discrimination with likelihood-ratios. Ann. Math. Statist. 42 1339-1347.

Bhattacharya, P. K. and Smith, R. P. (1972). SPRT for the mean value function of a Gaussian process. To be published in Ann. Math. Statist.

Kinderman, A. (1972). See Ref. 4.

4. General Theory of Classification When the Information About the Distribution is Based on Samples.

When the class-distributions $F_1, F_2, \ldots, F_m$ (and the mixture probabilities $\xi_1, \ldots, \xi_m$, in case of a mixed population) are not completely known, information on them is available through a training sample (TS). As described in Section 2, a training sample may be obtained separately from each π_i $(i = 1, \ldots, m)$, or from a mixture of these populations and, in that case, the sample may be supervised, unsupervised or post-supervised. Let n be the total size of the training sample and n_i be the size of the sample from π_i. Let n_0 be the size of the sample from π_0 which has to be classified and we shall denote such a sample by "CS."

Classification rules are generally devised using the following methods:

(a) Plug-in rules: Under complete knowledge of F_i's (and ξ_i's, in case of mixed population) a good rule (e.g., Bayes rule, minimax rule, LR rule, MD rule etc.) δ is chosen. A plug-in rule $\hat{\delta}$ is obtained by replacing the F_i's (and ξ_i's) in δ by the corresponding estimates obtained from TS. When δ involves only the class densities or the parameters (in the parametric case) the corresponding estimates of the densities or the parameters are used in $\hat{\delta}$. In case of a MD rule using a distance function d, estimates of the distributions have to be chosen appropriately so that d is defined for these estimates.

(b) LR rules and ML rule: Suppose the class-densities are known except for some parameters. Let $L(TS)$ denote the likelihood of the training sample and $L_i(CS)$ denote the likelihood of CS under the hypothesis $\pi_0 = \pi_i$. Let $\lambda_i = \sup[L_i(CS)L(TS)]$, the supremum being taken over the parametric space. A LR rule classifies CS into π_i, iff

$$k_i\lambda_i = \max_{1\le j\le m} [k_j\lambda_j],$$

where k_i's are non-negative constants; ties may be resolved in some manner. A ML rule is a LR rule with equal k_i's. The concept of a LR rule for the classification problem is due to Anderson (1951).

(c) Best-of-class or constructive rules: Such a rule is given by the one which optimizes certain specified criteria in a given class. One may consider Bayes rules, admissible rules, minimax rules or characterize a (reasonable) complete class following the general theory of statistical decision functions; the class of rules may be restricted by some invariance requirement. Note that in this case the action space of the statistician is finite. See Wald (1950), Kudo (1959, 1960). Another possibility is to consider some criteria depending on PMC's and use Neyman-Pearson theory and its extensions. See Rao (1954).

The criteria may be defined "empirically" as follows. Suppose the criteria for evaluating the performance of a rule is given by a real-valued function (e.g., PCC). For each rule, an estimate of the value of the function corresponding to the rule is defined in terms of TS. Then an empirical best-of-class rule is the one for which such an estimate is the maximum in a given class of rules. See Glick (1969).

The main problem in obtaining plug-in rules is to get reasonable estimates of the distributions (or the densities, or the parameters). Generally, maximum-likelihood or some other consistent estimates are used. For estimation, especially in the non-supervised case, there is a huge literature and it is not possible to discuss these papers in this review. For an early work on non-supervised estimation, see Pearson (1894). For estimation by potential function method and stochastic approximation method, especially in the non-supervised case, see Fu (1968) and Patrick (1972) and references therein.

General results on asymptotic properties of plug-in rules are given in Hoel and Peterson (1949), Fix and Hodges (1951), Das Gupta (1964), Van Ryzin (1966), Bunke (1967), Glick (1969, 1972).

See Sections 5, 6, 7 and Glick (1969, 1972), in particular, for studies on estimation of PMC's of a rule.

There is no systematic work on sequential rules. See Fu (1968), Patrick (1972) and Kurz and Woinsky (1969) in this connection. Kinderman (1972) studied some sequential rules based on distance functions and suggested some rules based on the idea of "tests with power 1" of H. Robbins.

Classifiability.

Following the work of Hoeffding and Wolfowitz (Ann. Math. Statist., 1958) on distinguishability of distributions, Das Gupta and Kinderman (1972) (see also Kinderman (1972)) introduced an important notion termed as "classifiability." Suppose $F_1 \times F_2 \times \ldots \times F_m$ belongs to a certain set Ω of distributions. The set Ω is said to be classifiable finitely or sequentially if the PMC's can be controlled (arbitrarily) by some fixed sample-size rule or sequential rule, respectively, based on observations from $\pi_0, \pi_1, \ldots, \pi_m$. Different conditions are obtained for Ω to be classifiable. The structure of Ω is studied for resolving the problem whether observations from all the populations or some (specific) of them are required (or sufficient) so as to get a rule when the PMC's are controlled (arbitrarily).

Compound-Decision and Empirical Bayes Approaches.

Let $X_1, X_2, \ldots, X_n$ be independent random variables. The distribution of X_i is given by $F(\cdot, \theta_i)$ where $\theta_i \in [1,2,\ldots, m]$ and for each i the problem is to choose one of the decisions $\theta_i = j (j = 1,\ldots, m)$; this will be called the i^{th} component problem. The main bulk of the literature concerns with $m = 2$ and we shall only describe this case. The general theory is given in Robbins (1951). Let $T_n = (t_1,\ldots, t_n)$ be a decision rule, where t_i and $1 - t_i$ are the conditional probabilities of deciding $\theta_i = 1$ and 2, respectively, given the observations. Let R_n be the minimum value of the average of risks

for the n component problem when one considers only fixed rules given by $t_i = t(x_i)$. It is shown in Robbins (1951), Hannan and Robbins (1955) that there exists $\{T_n\}$ with $T_i = t_i(x_1, \ldots, x_n)$ such that for large n the risk of T_n is uniformly (in θ_i's) close to R_n; it is assumed that $F(\cdot, j)$ are known for all j. Hannan and Van Ryzin (1965) studied the rate of convergence of the risks of the above rules. Assuming that $X_1, X_2, \ldots, X_n$ occur in a sequence, Samuel considered "sequential" rules T_n with $t_i = t_i(x_1, \ldots, x_i)$. She (1963a) first characterized the minimal complete class under complete knowledge and in (1963b) proved a result similar to Hannan and Robbins restricting to "sequential" rules. For a nonparametric method, see Johns (1961).

The empirical Bayes method, suggested by Robbins (1964), was used by Hudimoto (1968) in devising a rule for classifying observations from the mixed population π_0. When the class-distributions are unknown, they are estimated (by nonparametric method) from a supervised TS. The risk of such a rule was also studied. Based on n_0 observations from a mixed distribution given by $\sum_{j=1}^{m} \xi_j F(x; \theta_j)$, Choi (1969) suggested to estimate the ξ_j's and the θ_j's by minimizing

$$\int [\Sigma \xi_j F(x; \theta_j) - \hat{F}(x)]^2 d\hat{F}$$

where $\hat{F}$ is the empirical c.d.f. These estimates are used to obtain a plug-in rule $\hat{\delta}$ from the Bayes rule δ. The asymptotic behavior of the conditional risk of $\hat{\delta}$ given the observations was studied.

See Tanaka (1970) for a method of approximating the difference of two posterior probabilities at the n^{th} stage for classifying a sequence of observations, each into one of two distributions.

References (4).

Pearson, Karl (1894). Contributions to the mathematical theory of evolution. Phil. Trans. Roy. Soc., London 185 71-110.

Hoel, P. G. and Peterson, R. P. (1949). A solution to the problem of optimum classification. Ann. Math. Statist. 20 433-438.

Wald (1950). See Ref. 1.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision functions. Proc. II Berkeley Symposium on Probability and Statistics. Univ. of California Press, 131-148.

Anderson, T. W. (1951). See Ref. 5.

Fix, E. and Hodges, J. L. (1951). Nonparametric discrimination: consistency properties. U.S. Air Force School of Aviation Medicine, Report No. 4. Randolph Field, Texas.

Rao, C. R. (1954). A general theory of discrimination when the information about alternative population is based on samples. Ann. Math. Statist. 25 651-670.

Kudo, A. (1959, 1960). See Ref. 5.

Johns, M. V. (1961). See Ref. 7.

Samuel, E. (1963a). Note on a sequential classification problem. Ann. Math. Statist. 34 1095-1097.

Samuel, E. (1963b). Asymptotic solutions of the sequential compound decision problem. Ann. Math. Statist. 34 1079-1094.

Robbins, H. (1964). The empirical Bayes approach to statistical decision functions. Ann. Math. Statist. 35 1-20.

Das Gupta, S. (1964). See Ref. 7.

Hannan, J. F. and Van Ryzin, J. (1965). Rate of convergence in the compound decision problems for two completely specified distributions. Ann. Math. Statist. 36 1743-1752.

Van Ryzin, J. (1966). See Ref. 7.

Bunke, O. (1966). Uber optimale verfahren der diskriminanzanalyse. Abhandlg. Deutsch. Akad. Wiss., Klasse Math., Phys. Tech. 5 35-41. (MR32-6624)

Hudimoto, H. (1968). On the empirical Bayes procedure (1). Ann. Inst. Statist. Math. 20 169-185.

Fu, K. S. (1968). See Ref. 1 (books).

Kurz, L. and Woinsky, M. M. (1969). See Ref. 7.

Choi, K. (1969). Empirical Bayes procedure for (pattern) classification with stochastic learning. Ann. Inst. Statist. Math. 21 117-125.

Glick, N. (1969). Estimating unconditional probabilities of correct classification. Stanford Univ. Department of Statistics, Tech. Report No. 3.

Tanaka, K. (1970). On the pattern classification problems by learning I, II. Bull. Math. Statist. (Japan) 14 31-50, 61-74.

Kinderman, A. (1972). On some properties of classification: classifiability, asymptotic relative efficiency, and a complete class theorem. Univ. of Minnesota, Department of Statistics Tech. Report No. 178.

Das Gupta, S. and Kinderman, A. (1972). On classifiability and requirements of different training samples. Unpublished.

Patrick, E. A. (1972). See Ref. 1 (books).

Glick, N. (1972). Sample-based classification procedures derived from density estimators. Jour. Amer. Statist. Assoc. 67 116-122.

5. Classification Into Multivariate Normal Populations--Nonsequential Methods.

A. Classification Into Two Multivariate Normal Distributions With the Same Covariance Matrix.

The distribution of X in π_i is $N_p(\mu_i, \Sigma)$, $i = 1, 2$. Suppose all the parameters μ_1, μ_2, and Σ are known and $X \sim N_p(\mu, \Sigma)$, where $\mu = \theta\mu_1 + (1-\theta)\mu_2$; for the classification problem $\theta = 0$ or 1. It is easily seen that $(\mu_1 - \mu_2)'\Sigma^{-1}X$ is a sufficient statistic for θ. The class of Bayes rules is the same as the class of LR rules. Typically, a LR rule δ_c classifies X into $N_p(\mu_1, \Sigma)$, iff

$$T(x) \equiv T(x; \mu_1, \mu_2, \Sigma) \equiv \|x-\mu_1\|_\Sigma^2 - \|x-\mu_2\|_\Sigma^2 \leq c,$$

where $\|a-b\|_\Sigma^2 = (a-b)'\Sigma^{-1}(a-b)$. The minimax Bayes rule (0-1 loss) is given by δ_0; it is also called the minimum distance (MD) rule (for Mahalanobis distance). The PMC of δ_c is given by

$$\alpha_1(\delta_c) = \Phi(-\frac{c+\Delta^2}{2\Delta}), \quad \alpha_2(\delta_c) = \Phi(\frac{c-\Delta^2}{2\Delta}),$$

where $\Delta^2 = \|\mu_1 - \mu_2\|_\Sigma^2$, and Φ is the c.d.f. of $N(0, 1)$. This classical case is treated in many papers and books, of which Welch (1939), Wald (1944), Rao (1952) and Anderson (1958) are worth mentioning. Recall that Fisher's LDF (in the population) is given by $(\mu_1 - \mu_2)'\Sigma^{-1}x$ which maximizes $[a'(\mu_1 - \mu_2)]^2/a'\Sigma a$ among all vectors a. Penrose (1947) suggested to consider the best LDF in terms of two linear functions of X given by the sum and a linear contrast of the components of X expressed in terms of their standard deviations; he called them the "size" and the"shape" respectively. He discussed the case when all the correlations are equal.

If the unknown parameters are structured in a special way, reasonable rules based on X can be found. For instance, Rao (1966) considered the

following structure, relevant for growth models: $\mu_i = \nu_i + \beta\theta_i$ $(i = 1,2)$, where ν_i, β are known but the vectors θ_i are unknown. By restricting to similar devisions of the sample space or by considering ancillary statistics, the problem is reduced to finding the usual LDF in terms of the projection of X on a space orthogonal to the column-space of β. This problem was originally posed by Burnaby (1966). Rao also treated the case when the covariance matrices are different.

Cochran (1962, 1964) studied the effects of the different components of X on Δ^2 (which determines the PMC of a LR rule), especially when all the correlations are equal.

When all the parameters are not known, random samples of sizes n_1 and n_2 from $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$ are used to get information on the parameters. (Sampling is different in the mixed-population case.) The literature in this area spans over (i) suggestions of some heuristic rules, especially the plug-in LR rules, (ii) distributions of classification statistics and expressions for PMC, (iii) estimation of the PMC of a given rule, and (iv) derivation of constructive rules.

The rules considered in the literature are usually of the type involving a classification statistic Z and a cut-off point c (i.e., classifies into π_1, iff $Z < c$), where Z is a function of X, $\bar{X}_1$, $\bar{X}_2$, and S, where $\bar{X}_1$ and $\bar{X}_2$ are the sample mean vectors and S is the sample pooled covariance matrix (the divisor being $n_1 + n_2 - 2 \equiv r$). The plug-in version of δ_c, denoted by $\hat{\delta}_c$, is based on the statistic $W \equiv T(X; \bar{X}_1, \bar{X}_2, S)$, when all the parameters are unknown. This statistic was proposed by Anderson (1951). More generally, one may consider a plug-in LR rule by replacing the unknown parameters in T by their respective estimates. Fisher (1936) and Wald (1943) suggested the plug-in LDF as the classification statistic, which is given by $U \equiv (\bar{X}_2 - \bar{X}_1)'S^{-1}X$. Anderson (1951, 1958) proposed the LR rules which have the following classification statistic:

$$[r + (1+1/n_1)^{-1}\|X-\bar{X}_1\|_S^2][r + (1+1/n_2)^{-1}\|X-\bar{X}_2\|_S^2]^{-1}.$$

When the cut-off point c is 1, this rule reduces to

$$V \equiv (1+1/n_1)^{-1}\|X-\bar{X}_1\|_S^2 - (1+1/n_2)^{-1}\|X-\bar{X}_2\|_S^2 \lesseqgtr 0 .$$

This is the same as $\hat{\delta}_0$, when $n_1 = n_2$. For known Σ, the LR rules involves the statistic V with S replaced by Σ. In the sequel, the rule $V \lesseqgtr 0$ will be called the ML rule and $\hat{\delta}_0$ will be called the MD rule; we shall use the same terminology when some of the known parameters are used instead of their estimates. Rao (1954) derived some rules restricting to invariance and local optimal conditions; the classification statistic for his rule (Σ unknown) will be called R. Matusita's (1967) minimum distance rules (for his distance function, see Sec. 5B) reduces to MD rules in this case; Matusita also considered the case when there are n_o observations to be classified and obtained some lower bounds for the PCC of the MD rule.

Rao (1946) suggested to test the hypothesis $\mu = (\mu_1 + \mu_2)/2$ by Hotelling's T^2-test and use the MD rule when this test is significant. Brown (1947) considered a problem where $\mu_i = \alpha + \beta w_i$ $(i = 1, 2)$, w_i being the classificatory variable (e.g., age). From training sample, α and β are estimated and using these estimates w is estimated for the observation to be classified; Brown extended this to more than 2 populations. Cochran (1964) posed the problem when the last q components of X have the same means in π_1 and π_2 and suggested to consider a statistic W^* (similar in form to W) in terms of the residuals in the first $p-q$ components of X after eliminating their (linear) regression on the last q components. Each of the statistics, U, V, W and R can be expressed as a linear function of the elements of a 2×2 random matrix

$$M = [m_{ij}] = (Y_1\ Y_2)'A^{-1}(Y_1\ Y_2),$$

where Y_1, Y_2 are independent $N_p(\cdot, I_p)$ vectors, and $A \sim W_p(n_1 + n_2 - 2; I_p)$ independently of Y_1 and Y_2. In particular, for V, W and R the means of Y_1 and Y_2 are proportional, and, moreover, V is a constant multiple of m_{12}. Wald (1943) gave a canonical representation of U, and Harter (1951) derived its distribution when p = 1. Sitgreaves (1952) derived the distribution of M when the means of Y_1 and Y_2 are proportional; Kabe (1963) derived it without this restriction. Bowker (1960) showed that W can be represented as a (rational) function of two independent 2×2 Wishart matrices one of which is noncentral. Bowker and Sitgreaves (1961) used this representation to find an asymptotic expansion of the c.d.f. of W in terms of n_1^{-1} and Hermite polynomials, when $n_1 = n_2$. Sitgreaves (1961) derived the distribution of m_{12} and explicitly obtained the PMC of the MD rule when $n_1 = n_2$. Elfving (1961) obtained an approximation to the c.d.f. of W for large $n_1 = n_2$ and p = 1. In the univariate case, Linhart (1961) gave an asymptotic expansion for the average PMC of the MD rule in powers of $(n_1 + n_2)/n_1 n_2$ and Hermite polynomials in Δ^2. Teichroew and Sitgreaves (1961) used an empirical sampling plan to obtain an estimate of the c.d.f. of W. Okamoto (1963) considered the statistic W where the degrees of freedom r of S is not necessarily $n_1 + n_2 - 2$, and gave asymptotic expansions of

$$P[(W-\Delta^2/2)/\Delta < k|\pi_1] \quad \text{and} \quad P[(W+\Delta^2/2)/\Delta < k|\pi_2]$$

in terms of n_1^{-1}, n_2^{-1} and r^{-1} as n_1, n_2, and r tend to ∞ and n_1/n_2 tend to a finite positive constant. Anderson (1972) obtained asymptotic expansions of the above probabilities with Δ^2 replaced by $D^2 \equiv \|\bar{X}_1 - \bar{X}_2\|_S^2$. Memon and Okamoto (1971) obtained an asymptotic expression for the c.d.f. of $(V+\Delta^2)/2\Delta$, when $\mu = \mu_1$.

Cochran (1964) numerically compared the PMC's (computed from Okamoto-expansion) of the rules $W^* \lesseqgtr 0$ with those of $W \lesseqgtr 0$ when $n_1 = n_2$ is large. Memon and Okamoto (1970) derived an asymptotic expansion for the distribution of W^* and the PMC of the W^*-rule in terms of n_1^{-1}, n_2^{-1} and r^{-1}.

John (1959, 1960) derived the distributions of the statistics U, V, W and Rao's statistic (when Σ is known), S being replaced by Σ and obtained explicitly the PMC when the cut-off point is 0. Some bounds for the PCC were also given by John. When Σ is unknown, and S is used for Σ, some approximations are given for the distributions and the PCC's.

For $p = 1$, $\mu_1 < \mu_2$, Friedman (1965) considered a rule: $X \lesseqgtr \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$ and compared its PMC with that of the rule $X \leq \frac{\mu_1 + \mu_2}{2}$, with approximations for large size of the training sample.

Recently, Das Gupta (1972) proved that for a large class of rules (including the MD and the ML rules--Σ known and Σ unknown) the PCC's are monotonic increasing functions of Δ^2.

Let $\delta \equiv \delta(\cdot; \mu_1, \mu_2, \Sigma)$ be a decision rule when all the parameters are known. We shall denote the plug-in version of δ by $\hat{\delta}$ by replacing the unknown parameters by their respective (standard) estimates. The conditional error probabilities of $\hat{\delta}$, given $\bar{X}_1$, $\bar{X}_2$ and S, are given by

$$e_i(\hat{\delta}) = P[\hat{\delta} \text{ classifies } X \text{ into } \pi_j | \bar{X}_1, \bar{X}_2, S; \mu = \mu_i],$$

$i \neq j$; $i, j = 1, 2$. The unconditional error-probabilities of $\hat{\delta}$ are $\alpha_i(\hat{\delta}) = E(e_i(\hat{\delta}))$. An estimate of $e_i(\hat{\delta})$ is given by $\hat{e}_i(\hat{\delta})$ which is obtained by replacing the unknown parameters in $e_i(\hat{\delta})$ by their standard estimates. Similarly $\hat{\alpha}_i(\delta)$ and $\hat{\alpha}_i(\hat{\delta})$ are defined.

In the literature, the error-probabilities of the minimax rule δ_0 (parameters known) and its plug-in version $\hat{\delta}_0$ (the MD rule) are mostly considered. When

Σ is known, John (1961) derived the distributions of $e_i(\hat{\delta}_0)$ and obtained their means; similar results are obtained when the cut-off point is not 0 and only approximations are given when Σ is unknown and n_1 and n_2 are large. In (1964) John considered the similar problem except that μ may be different from μ_1 or μ_2. John (1963) studied the conditional PMC's of the rules defined by the classification statistics $(1 + \frac{1}{n_1})^{-1}\|X-\bar{X}_1\|_\Sigma^2 - \eta(1 + \frac{1}{n_2})^{-1}\|X-\bar{X}_2\|_\Sigma^2$ and Rao's statistic R, when Σ is known. Dunn and Varady (1966) empirically studied (Monte Carlo methods) $1 - \hat{\alpha}_i(\delta_0)$, $1 - e_i(\hat{\delta}_0)$ and $1 - \hat{e}_i(\hat{\delta}_0)$ and derived a confidence interval for the conditional error probabilities of $\hat{\delta}_0$. Geisser (1967) considered a prior measure for the paramters whose (improper) density is proportional to $|\Sigma|^{(p+1)/2}$. Using the posterior distribution of the parameters (given $\bar{X}_1$, $\bar{X}_2$ and S) he obtained confidence bounds for $e_i(\hat{\delta}_0)$; for large n_1, n_2 he used normal approximations. Several estimates of $e_1(\hat{\delta}_0)$, $\alpha_1(\hat{\delta}_0)$, $\alpha_1(\delta_0)$ are suggested in the literature of which the following are of main types: (i) Smith's (1947) reallocation or counting estimates, (ii) Lachenbruch's (1967) deletion-counting estimate, (iii) Fisher's estimate $\hat{e}_1(\hat{\delta}_0) = \Phi(-D/2)$ or the estimate obtained by replacing Δ in $\Phi(-\Delta/2)$ by some other estimate, (iv) the leading term in the Okamoto-expansion and replacing Δ^2 by its estimate, (v) estimates obtained from additional training sample. It follows from Hills (1966) that $\alpha_i(\hat{\delta}_0) > \alpha_i(\delta_0)$ when $n_1 = n_2$. For $p = 1$, Hills (1966) obtained the distribution of $\hat{e}_1(\hat{\delta}_0)$ and compared the expectations of $e_1(\hat{\delta}_0)$, $\hat{e}_1(\hat{\delta}_0)$ and those of the counting estimate by exact expressions and numerical computations. In 1967, Lachenbruch proposed the deletion-counting method for estimation. Lachenbruch and Mickey (1968) suggested some estimates of Δ^2 and studied empirically the behavior of the estimates (i)-(iv). Brofitt (1969) derived the uniformly minimum variance estimates of the mean values of Smith's and Lachenbruch's estimates and suggested some other estimators with smaller mean-square errors. Sorum (1971) obtained

some estimates based on additional observations. For known Σ, she derived the means, the variances and approximation to the mean-square errors of most of the estimates and studied these estimates numerically when Σ is unknown (1972a, 1972b). Dunn (1971) studied the average PCC of $\hat{\delta}_0$ and Lachenbruch's estimates (using his estimate of Δ^2) for $n_1 = n_2$ by Monte Carlo methods. For $p = 1$, Sedransk and Okamoto (1971) obtained asymptotic expansions for the mean-square errors of several estimates. Recently, Das Gupta (1972) obtained some results on Fisher's and Smith's estimates which generalize Hill's (1966) results.

Chan and Dunn (1972) studied the effect of missing data on the PMC of $\hat{\delta}_0$ by Monte Carlo methods using several standard techniques of handling missing data. Srivastava and Zaatar (1972) derived the ML rule when Σ is known and the samples from the two populations are incomplete (all the p components are not available on each unit sampled) and showed that this rule is admissible Bayes. Lachenbruch (1966) posed the problem when the parent populations of the observations in the training sample are incorrectly identified. Mclachan (1972) derived asymptotic expressions for the mean and the variance of $e_i(\hat{\delta}_0)$ incorporating the possibility of incorrect identification of the training sample.

Following Glick (1972) it can be shown that as $n_1, n_2 \to \infty$, $\hat{\alpha}_i(\delta) \to \alpha_i(\delta)$ a.s. uniformly in the class of all rules (not based on training data). Furthermore, if δ is a LR rule, then $\alpha_i(\hat{\delta}) \to \alpha_i(\delta)$ a.s. and $\hat{\alpha}_i(\hat{\delta}) \to \alpha_i(\delta)$. For related results, see Glick (1969, 1972) and for slightly weaker results see Fix and Hodges (1950), Bunke (1964). Kinderman (1972) suggested a measure of the relative asymptotic efficiencies of two rules by the limit of the ratio of minimum total sample sizes required by the two rules to achieve a maximum probability of error α, as $\alpha \to 0$. In particular, he illustrated this concept

by comparing a two-sample rule based on samples from π_0 and π_1 and a three-sample rule using Anderson's statistic when the populations are univariate normal with variance 1 and $\Delta = |\mu_1 - \mu_2| > 0$.

There are many ad hoc methods for choosing "good" components of the vector X. Cochran (1961, 1964) studied the effect of the different components of X on Δ^2, especially when all the correlations are equal. Urbakh (1971) made a similar study on Δ^2, as well as, on Lachenbruch's estimate of Δ^2. Linhart (1961) made a numerical comparison of the effectiveness of selecting components by $\Phi(-\Delta/2)$ and the average PMC of $\hat{\delta}_0$. Weiner and Dunn (1966) also studied empirically three methods for selecting components.

In the normal case, Glick (1969) obtained some interesting results for the 'best-of-class ' rules. Let C_{LD} be the class of all rules based on linear (discriminants) functions of X (i.e., partitioning the sample space into two half spaces). Let δ^* be a rule in C_{LD} which maximizes (in C_{LD}) the average (over some known prior or the standard estimates of the proportions in the mixture) of the proportions of the training sample correctly classified. Then this maximum value converges (a.s.) to the PCC of the best (Bayes) rule and the risk of δ^* converges a.s. to the Bayes risk as the sample size in the training sample increases to ∞.

When the training sample comes from a mixed population different methods are available to estimate the parameters and the proportions in the mixture, if they are unknown. For the supervised case, there is not much change in the theory and the methods from the usual case discussed before. For some asymptotic results see Glick (1969, 1972). In the non-supervised case, there is a good deal of literature; for this and relevant references, see Fu (1968), Patrick (1972); for an earlier work see Pearson (1894).

Rao (1954) derived an optimal rule in the class of rules for which the probabilities of error depend only on Δ^* using the following criteria: (i) to

minimize a linear combination of the derivatives of the error-probabilities with respect to Δ at $\Delta = 0$ subject to the condition that the error-probabilities at $\Delta = 0$ bear a given ratio. (ii) The above criterion with the additional restriction that the derivatives of the error-probabilities at $\Delta = 0$ bear a given ratio. Rao separately treated the problem according as Σ is known or unknown. When Σ is known, Kudo (1959, 1960) showed that the ML rule has the maximum PCC among all translation-invariant rules δ for which the error-probabilities depend on Δ^2, and

$$\alpha_1(\delta;\ \Delta^2 = \Delta_1^2) = \alpha_2(\delta;\ \Delta^2 = \Delta_2^2)$$

for all Δ_1 and Δ_2 such that $(1 + 1/n_1)^{-1}\Delta_1^2 = (1 + 1/n_2)^{-1}\Delta_2^2$. He also showed that this rule is most stringent in the above class without the requirement of translation-invariance. When Σ is known, Ellison (1962) obtained a class of admissible Bayes rules which includes the MD and ML rules. In this case, Das Gupta (1962, 1965) showed that the ML rule is admissible Bayes (with a different prior and a general loss function) and minimax (unique minimax under some mild conditions). When Σ is unknown, similar results were obtained by Das Gupta (1962, 1965), restricting to the class of rules invariant under translation and the full linear group. For $p = 1$, $n_1 = n_2$, Bhattacharya and Das Gupta (1964) obtained a class of Bayes rules and showed that the MD rule is minimax Bayes. Srivastava (1964) also obtained a class of Bayes rules when Σ is unknown. Geisser (1964) used a prior (improper) density which is proportional to $|\Sigma|^{v/2}$, $v \leq n_1 + n_2$ and $v = 0$ when Σ is known; he derived the (improper) Bayes rules for these priors which are the likelihood-ratio rules in respective cases. For similar analysis, see Geisser (1966). Kiefer and Schwartz (1965) indicated a method to obtain a broad class of Bayes rules which are admissible; in particular, they showed that the LR rules are admissible Bayes when Σ is unknown and $r + 1 > p$. Marshall and Olkin (1968) derived Bayes rules for normal distributions in their special set-up. When $p = 1$, $n_1 = n_2$ and the number

of observations to be classified is $n(\geq 1)$, Kinderman (1972) characterized an essentially complete class of rules, invariant under translation and change of signs.

B. Classification Into Two Multivariate Normal Populations With Different Covariance Matrices.

The distribution of X in π_i is taken as $N_p(\mu_i, \Sigma_i)$, $i = 0,1,2$; furthermore, it is known that Σ_1 and Σ_2 are different. Generally, three cases are considered: (i) $(\mu_0, \Sigma_0) = (\mu_i, \Sigma_i)$ for some $i = 1, 2$, (ii) $\mu_0 = \mu_i$, for some $i = 1, 2$, (iii) $\Sigma_0 = \Sigma_i$ for some $i = 1, 2$.

When the parameters are known, the LR statistic was studied by Cavalli (1945) ($p = 1$), Smith (1947), Okamoto (1963) ($\mu_0 = \mu_1 = \mu_2$), Cooper (1963, 1965), Bartlett and Please (1963) ($\mu_0 = \mu_1 = \mu_2 = 0$, $\Sigma_i = (1-\rho_i)I_p + \rho_i J_p$, $i = 1, 2$), Bunke (1964), Han (1968) ($\Sigma_i = (1-\rho_i)I_p + \rho_i J_p$, $i = 1, 2$), Hann (1969) ($\Sigma_1 = d\Sigma_2$, $d > 1$), Han (1970) (Σ_i's are of circular type).

Kullback (1952, 1958) suggested a rule based on the linear statistic which maximizes the divergence $J(1, 2)$ between $N_p(\mu_1, \Sigma_1)$ and $N_p(\mu_2, \Sigma_2)$. He also obtained some partial results on deriving the optimum class of rules based on linear functions of X from Neyman-Pearson viewpoint (i.e., minimizing one PMC by controlling the other). Clunies-Ross and Riffenburgh (1960) studied this problem geometrically. Anderson and Bahadur (1962) derived the minimax rule and characterized the minimal complete class after restricting to the class of rules based on linear functions of X. Banerjee and Marcus (1965) studied the form of this minimax rule.

Gilbert (1969) derived the PMC of a LR rule when the parameters are known and compared it with the PMC of the corresponding LR rule when $\Sigma_1 = \Sigma_2$. For the later rule he obtained the optimum cut-off point for which the total PMC is minimized.

Lbov (1964) studied the PMC when p is large and the parameters are known. See Grenander (1972) for a similar problem.

Anderson (1964) studied the problem of choosing components by minimizing Bayes risk when the distributions are univariate normal.

When μ_0 equals either μ_1 or μ_2, and the covariance matrices are known, a class of admissible Bayes rules was obtained by Ellison (1962); in particular he showed that the MD and the ML rules are admissible Bayes.

Okamoto (1963) derived the minimax rule and the form of a Bayes rule when the parameters are known; he studied some properties of the Bayes' risk function, and suggested a method for choosing components. He also treated the case when Σ_i's are unknown, and the common value of μ_i's may be known or unknown. The asymptotic distribution of the plug-in log(LR) statistic was also obtained by Okamoto. Bunke (1964) derived the minimax rule and the form of a Bayes rule and proved that the plug-in minimax rule is consistent. Following the method of Kiefer and Schwartz (1965), Nishida (1971) obtained a class of admissible Bayes rules when the parameters are unknown.

Matusita (1967) considered a minimum distance rule and suggested its plug-in version by replacing the unknown parameters by their respective estimates; the distance between two distributions with p.d.f.'s p_1 and p_2 with respect to a σ-finite measure m was taken as

$$[\int(\sqrt{p_1(x)} - \sqrt{p_2(x)})^2 dm]^{\frac{1}{2}} .$$

He separately treated the different cases according as the μ_i's and Σ_i's are known or unknown, and obtained some bounds for the PCC.

When $\Sigma_1 = d\Sigma_2$ $(d > 1)$, the distributions of the log(LR) statistic and its plug-in version (by replacing the mean vectors by their estimates) were derived by Han (1969). Similar results were obtained by Han (1970) when Σ_i's are of "circular" type.

Chaadha and Marcus (1968) studied (mainly simulation) the behavior of some estimates of a measure of divergence defined as $2(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)$.

Ayoma (1950) considered rules of the form $X \lesseqgtr x_0$ and found the optimum value of x_0 which minimizes the PCC when X is a mixture of two univariate normal distributions; the mixture ratio may be known or unknown.

C. Classification Into More Than Two Multivariate Normal Populations.

The distribution of X in π_i is taken as $N_p(\mu_i, \Sigma_i)$, $i = 0,1,\ldots, m$. In most of the cases the results for $m = 2$ are extended in a straightforward way, and discussions on this case may be found in many papers cited in Sections 5A and 5B. In particular, see Fisher (1938), Day and Sandomire (1942), Brown (1947), Rao (1952), Anderson (1958), Rao (1963). Generalizing Fisher's LDF, one considers the eigenvectors of the "between means" matrix in the metric of "within error" matrix. For other criteria, see Uematu (1964).

Das Gupta (1962) considered the problem when $\mu_1,\ldots, \mu_m$ are linearly restricted (as in the linear model in MANOVA) and showed that the ML rule is admissible Bayes when the common covariance is known. Following Kiefer and Schwarz (1965), Srivastava (1967) obtained similar results when the common covariance matrix is unknown.

Cacoullos (1965) considered the case when the distribution of X in π_i is $N_p(\mu_i,\Sigma)$, $i = 0, 1,\ldots,m$, and μ_0 is not necessarily equal to one of $\mu_1,\ldots,\mu_m$; the problem is to choose a π_i which is nearest to π_0 (in the sense of Mahalanobis-distance). When $\mu_1,\ldots,\mu_m$, and Σ are known he obtained a unique invariant minimax rule allowing for indifference regions; for the m-decision problem he obtained a class of admissible Bayes rules including the minimax rule. In a later paper (1965), Cacoullos obtained a class of Bayes rules when Σ and $\mu_1,\ldots,\mu_m$ are unknown.

References (5)

Fisher, R. A. (1936). Ann. Eugen. See Ref. 1.

Fisher, R. A. (1938). Ann. Eugen. See Ref. 1.

Welch, B. L. (1939). Biom. See Ref. 1.

Day, B. B. and Sandomire, M. M. (1942). J. Amer. Statist. Assoc. See Ref. 1.

Wald, A. (1944). Ann. Math. Statist. See Ref. 1.

Cavalli, L. L. (1945). Mem. Ist. Ital. Idrobiol. See Ref. 1.

Rao, C. R. (1946). Sankhyā. See Ref. 1.

Penrose, L. S. (1947). Ann. Eugen. See Ref. 1.

Smith, C. A. B. (1947). Ann. Eugen. See Ref. 1.

Brown, G. W. (1947). Ann. Math. Statist. See Ref. 1.

Rao, C. R. (1948). J. Roy. Statist. Soc. (B). See Ref. 1.

Rao, C. R. (1949b). Sankhyā. See Ref. 1.

Rao, C. R. (1950). Sankhyā. See Ref. 1.

Aoyama, H. (1950). Ann. Inst. Statist. Math. See Ref. 1.

Anderson, T. W. (1951). Classification by multivariate analysis. Psychometrika 16 631-650.

Harter, H. L. (1951). On the distribution of Wald's classification statistic. Ann. Math. Statist. 22 58-67.

Rao, C. R. (1952). Wiley. See Ref. 1 (books).

Sitgreaves, R. (1952). On the distribution of two random matrices used in classification procedures. Ann. Math. Statist. 23 263-270.

Kullback, S. (1952). An application to information theory to multivariate analysis. Ann. Math. Statist. 23 88-102.

Rao, C. R. (1954). A general theory of discrimination when the information on alternative hypotheses is based on samples. Ann. Math. Statist. 25 651-670.

Kudo, A. (1959). The classification problem viewed as a two-decision problem I. Mem. Fac. Sci. Kyushu Univ. Ser. A. 13 96-125.

Kullback, S. (1959). Information Theory and Statistics. Wiley, New York (1968, Dover, New York).

John, S. (1959). The distribution of Wald's classification statistic when the dispersion matrix is known. Sankhyā 21 371-376.

Kudo, A. (1960). The classification problem viewed as a two-decision problem II. Mem. Fac. Sci. Kyushu Univ. Ser. A. 14 63-83.

John, S. (1960). On some classification statistics I, II. Sankhyā 22 301-308, 309-316. (Correction: Sankhyā 23 (1961) 308.)

Bowker, A. H. (1960). A representation of Hotelling's T^2 and Anderson's classification statistic. Contrib. Probability and Statistics (Hotelling Vol.) 142-149.

Clunies-Ross, C. W. and Riffenburgh, R. H. (1960a). Linear discriminant analysis. Pacif. Sci. 14 251-256.

John, S. (1961). Errors in discrimination. Ann. Math. Statist. 32 1125-1144.

Sitgreaves, R. (1961). Some results on the distribution of the W-classification statistic. Stud. Item. Anal. Pred. Ed. H. Solomon. 241-251. Stanford Univ. Press, Stanford, California.

Bowker, A. H. and Sitgreaves, R. (1961). An asymptotic expansion for the distribution function of the W-classification statistic. Ibid. 285-292.

Bowker, A. H. (1961). A representation of Hotelling's T^2 and Anderson's classification statistic W in terms of simple statistic. Ibid. 285-292. (Reprint of Bowker (1960).)

Teichroew, D. and Sitgreaves, R. (1961). Computation of an empirical sampling distribution for the W-classification statistic. Ibid. 252-275.

Elfving, G. (1961). An expansion principle for distribution functions with applications to Student's t-statistic and the one-dimensional classification statistic. Ibid. 276-284.

Okamoto, M. (1961). Discrimination for variance matrices. Osaka Math. Jour. 13 1-39.

Linhart, H. (1961). Fur wahl von variablen in der Tremanalyse. Metrika 4 126-139.

Cochran, W. G. (1961). On the performance of the linear discriminant function. Bull. Intl. Statist. Inst. 34 436-446. (Reprinted in Technometrics 6 (1964) 179-190.)

Anderson, T. W. and Bahadur, R. (1962). Classification into two multivariate normal distributions with different covariance matrices. Ann. Math. Statist. 33 420-431.

Ellison, B. E. (1962). A classification problem in which information about alternative distributions is based on samples. Ann. Math. Statist. 33 213-223.

Das Gupta, S. (1962). On the optimum properties of some classification rules. Inst. Statist. Mimeo No. 333. Univ. of N.C., Chapel Hill. (See abstract: Ann. Math. Statist. 33 1504.)

Cooper, P. W. (1962). The hyperplane in pattern recognition. Cybernetica 5 215-238.

Cooper, P. W. (1962). The hypersphere in pattern recognition. Inform. Control 5 324-346.

John, S. (1963). On classification by the statistics R and Z. Ann. Inst. Statist. Math. 14 237-246. (Correction: Ann. Inst. Statist. Math. 17 (1965) 113.)

Rao, M. M. (1963). Discriminant Analysis. Ann. Inst. Statist. Math. 15 11-24.

Kabe, D. G. (1963). Some results on the distribution of two random matrices used in classification procedures. Ann. Math. Statist. 34 181-185.

Okamoto, M. (1963). An asymptotic expansion for the distribution of linear discriminant function. Ann. Math. Statist. 34 1286-1301. (Correction: Ann. Math. Statist. 39 (1968) 1358-1359.)

Bartlett, M. S. and Please, N. W. (1963). Discrimination in the case of zero mean differences. Biometrika 50 17-21.

Cooper, P. W. (1963). Statistical classification with quadratic forms. Biometrika 50 439-448.

John, S. (1964). Further results on classification by W. Sankhyā A 26 39-46.

Bhattacharya, P. K. and Das Gupta, S. (1964). Classification into exponential populations. Sankhyā A 26 17-24.

Bunke, O. (1964). Uber optimale verfahren der diskriminanzanalyse. Abl. Deutsch Akad. Wiss. Klasse Math Phys. Tech. 4 35-41 (MR32-6624).

Uematu, T. (1964). On a multidimensional linear discriminant function. Ann. Inst. Statist. Math. 16 431-437.

Anderson, T. W. (1964). On Bayes procedures for a problem with choice of observations. Ann. Math. Statist. 35 1128-1135.

Lbov, G. S. (1964). Errors in the classification of patterns for unequal covariance matrices. Akad. Nauk SSSR Sibirisk Otdel Inst. Mat. Vye. Sistemy 14 31-38 (MR31-5724).

Geisser, S. (1964). Posterior odds for multivariate normal classification. J. Roy. Statist. Soc. Ser. B 26 69-76.

Srivastava, M. S. (1964). Optimum procedures for classification and related topics. Stanford University Department of Statistics Tech. Report No. 11.

Cochran, W. G. (1964). Comparison of two methods of handling covariates in discriminatory analysis. Ann. Inst. Statist. Math. 16 43-53.

Das Gupta, S. (1965). Optimum classification rules for classification into two multivariate normal populations. Ann. Math. Statist. 36 1174-1184.

Kiefer, J. and Schwartz, R. (1965). Admissible Bayes character of T^2-, R^2- and other full invariant tests for classical multivariate normal problem. Ann. Math. Statist. 36 747-770.

Ellison, B. E. (1965). Multivariate normal classification with covariance known. Ann. Math. Statist. 36 1787-1793.

Banerjee, K. S. and Marcus, L. F. (1965). Bounds in a minimax classification procedure. Biometrika 52 653-654.

Cooper, P. W. (1965). Quadratic discriminant functions in pattern recognition. IEEE Trans. Inform. Theory IT-11 313-315.

Friedman, H. D. (1965). On the expected error in the probability of misclassification. Proc. IEEE 53 658-659.

Cacoullos, T. (1965). Comparing Mahalanobis distances. I,II. Sankhyā A 27 1-22,23-32.

Cacoullos, T. (1966). On a class of admissible partitions. Ann. Math. Statist. 37 189-95.

Weiner, J. M. and Dunn, O. J. (1966). Elimination of variates in linear discriminant analysis. Biometrics 22 268-275.

Geisser, S. (1966). Predictive discrimination. Proc. Internat. Symp. Multiv. Analysis. Ed. P. R. Krishnaiah. Academic Press, New York.

Cochran, W. G. (1966). Analyse des classifications d'ordre. Rev. Statist. Appl. 14 5-17.

Hills, M. (1966). Allocation rules and their error rates. J. Roy. Statist. Soc. Ser. B 28 1-31.

Burnaby, T. P. (1966). Growth invariant discriminant functions and generalized distances. Biometrics 22 96-110.

Rao, C. R. (1966). Discriminant function between composite hypothesis and related problems. Biometrika 53 339-345.

Lachenbruch, P. A. (1966). Discriminant analysis when the initial samples are misclassified. Technometrics 8 657-662.

Dunn, O. J. and Varady, P. V. (1966). Probabilities of correct classification in discrimination analysis. Biometrics 22 908-924.

Srivastava, M. S. (1967). Classification into multivariate normal populations when the population means are restricted. Ann. Inst. Statist. Math. 19 473-478.

Bunke, O. (1967). Z. Wahr Theor. und Verwindte Gebiete (MR35-6267). See Ref. 4.

Geisser, S. (1967). Estimation associated with linear discriminants. Ann. Math. Statist. 38 807-817.

Day, N. E. and Kerridge, D. F. (1967). A general maximum likelihood discriminant. Biometrics 23 313-323.

Matusita, K. (1967). Classification based on distance in multivariate Gaussian case. Proc. 5th Berk. Symp. Math. Stat. Prob. 1 299-304. Univ. of Calif. Press, Berkeley.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics 23 639-645.

Fu, K. S. (1968). Academic Press. See Ref. 1 (books).

Chaddha, R. L. and Marcus, L. F. (1968). An empirical comparison of distance statistics for populations with unequal covariance matrices. Biometrics 24 683-

Han, Chien Pai (1968). A note on discrimination in the case of unequal covariance matrices. Biometrika 55 586-587.

Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. Technometrics 10 1-11.

Marshall, A. W. and Olkin, I. (1968). J. Roy. Statist. Soc. Ser. B. See Ref. 3 .

Han, Chien Pai (1969). Distribution of discriminant function when covariance matrices are propositional. Ann. Math. Statist. 40 979-985.

Gilbert, E. S. (1969). The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. Biometrics 25 505-516.

Glick, N. (1969). Stanford Univ. Department of Statistics Tech. Report. See Ref. 4.

Brofitt, J. D. (1969). Estimating the probability of misclassification based on discriminant function techniques. PhD dissertation, Colorado State Univ., Fort Collins.

Han, Chien Pai (1970). Distribution of discriminant function in circular models. Ann. Inst. Statist. Math. 22 117-125.

Memon, A. Z. and Okamoto, M. (1970). The classification statistic W^* in covariate discriminant analysis. Ann. Math. Statist. 41 1491-1499.

Memon, A. Z. and Okamoto, M. (1971). Asymptotic expansion of the distribution of the Z-statistic in discriminant analysis. J. Multiv. Anal. 1 294-307.

Nishida, N. (1971). A note on the admissible tests and classifications in multivariate analysis. Hiroshima Math. J. 1 427-434.

Dunn, O. J. (1971). Some expected values for probabilities of correct classification in discriminant analysis. Technometrics 13 345-353.

Urbakh, V. Yu. (1971). Linear discriminant analysis: loss of discriminating power when a variate is omitted. Biometrics 27 531-534.

Sorum, M. J. (1971). Estimating the conditional probability of misclassification. Technometrics 13 333-343.

Sedransk, N. and Okamoto, M. (1971). Estimation of the probabilities of misclassification for a linear discriminant function in the univariate normal case. Ann. Inst. Statist. Math. 23 419-436.

Sorum, M. J. (1972a). Three probabilities of misclassification. Technometrics 14 309-316.

Sorum, M. J. (1972b). Estimating the expected and the optimal probabilities of misclassification. Technometrics 14 935-943.

Glick, N. (1972). J. Amer. Statist. Assoc. See Ref. 4.

Mclachan, G. J. (1972). Asymptotic results for discriminant analysis when the initial samples are misclassified. Technometrics 14 415-422.

Chan, L. S. and Dunn, O. J. (1972). The treatment of missing values in discriminant analysis-I. J. Amer. Statist. Assoc. 67 433-477.

Kinderman, A. (1972). Univ. of Minnesota, Department of Statistics Tech. Rep. No. 178. See Ref. 4.

Anderson, T. W. (1972). An asymptotic expansion of the distribution of the "Studentized" classification statistic W. Stanford Univ., Department of Statistics Tech. Report No. 9.

Anderson, T. W. (1972). Asymptotic evaluation of the probabilities of misclassification by linear discriminant functions. Stanford Univ., Department of Statistics Tech. Report No. 10.

Das Gupta, S. (1972). Probability inequalities and error in classification. Univ. of Minnesota, School of Statistics Tech. Report No. 190.

Srivastava, J. N. and Zaatar, M. K. (1972). On the maximum likelihood classification rule for incomplete multivariate and its admissibility. J. Multiv. Anal. 2 115-126.

Grenander, Ulf (1972). Asymptotic distribution of quadratic forms: large deviations. (Unpublished.)

Govindarajulu, Z. and Gupta, A. K. (1972). Some new classification rules for c univariate normal populations. Dept. of Statist., Univ. of Michigan, Tech. Report No. 14.

6. Discrete and Other Non-normal Distributions.

The papers are grouped according to the type of distribution and the nature of the problem considered. A short review is given for each paper.

(a) Multinomial distribution.

The random variable X is distributed as a multinomial distribution with k cells in each of the populations.

Matusita (1956):

A minimum distance rule is proposed based on samples of sizes n, n_1 and n_2 from the populations π, π_1 and π_2, respectively. The distance is computed for the sample c.d.f.'s and the distance function is taken as the square root of

$$\|F - G\|^2 = \sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2,$$

where $(p_1, \ldots, p_k)$ and $(q_1, \ldots, q_k)$ all cell probabilities corresponding to the distributions F and G, respectively. He obtained lower bounds for PCC and approximate value of the PCC when the sample sizes are large; the case $n = 1$ is also discussed.

Chernoff (1956):

The distribution of X in π_1 is the multinomial with equal cell-probabilities and a multinomial with unknown cell probabilities in π_2. A sample of size n_2 is available from π_2 and the problem considered is to classify a sample of size n_0 from π_0 into π_1 or π_2. Results are directed towards applications when the number of cells is large, n_0 and n_2 are large, and the ratio of error probabilities is either very large or very small. In a certain class, an 'optimal' rule is obtained which classifies into π_2 if the sum of the frequencies of all the cells for which the sample from π_2 provides non-zero frequencies is too large.

Wesler (1959):

The distribution of X in π_i is taken as a multinomial with cell probabilities being any permutation of a given probability vector $p^{(i)}$, $(i = 1, 2)$. The problem considered is to classify a sample of n_0 observations from π_0, by minimaxing one error probability when the maximum of the other error probability is held fixed. He obtained an approximate solution for large n_0 and considered the case $k = 2$.

Cochran and Hopkins (1961):

They obtained the form of the Bayes rules and considered, in particular, the 'maximum likelihood' rule. For this rule they discussed the effect of 'plug-in' on the PMC and suggested a correction for bias.

Raiffa (1961):

See Section 3. Multinomial distributions, and, in general, discrete distributions are included in the development of theories.

Hills (1966):

This paper contains some theoretical developments on the errors of misclassification for the 'ML' rule in the two-population case. In particular, it is shown that for $k = 2$ PMC for the 'ML' rule is greater than the corresponding PMC for the ML rule obtained under complete knowledge of distributions and Smith's reallocation estimate of the PMC underestimates the PMC of the ML rule. He obtained normal approximations for the expected value of the reallocation estimate, plug-in estimate of the PMC of the ML rule and its expected value. The effectiveness of these estimates are compared through a numerical study.

Bunke (1966):

For multinomial distributions, a property of the estimated (with empirical c.d.f.'s) minimax rule is studied from asymptotic viewpoint.

Glick (1969):

The development is for general discrete distributions but also specialized for multinomial distributions. This paper generalizes some of the results of Cochran and Hopkins (1961) and Hills (1966) and furnishes rigorous proofs. The sample space $\mathcal{X}$ of the random variable X is taken as $\mathcal{X} = (x_1, x_2, \ldots, x_k, \ldots)$. The population π_0 is considered as a mixture of the populations $\pi_1, \ldots, \pi_m$. The rule δ (Bayes) which maximizes the PCC is dealt with throughout. Let $\hat{\delta}$ be the plug-in version of δ using the "supervised training" data. Let γ = PCC for δ, $c(\hat{\delta})$ = the conditional (given the training data) PCC for $\hat{\delta}$, $\hat{c}(\hat{\delta})$ = plug-in version of $c(\hat{\delta})$. The following results are obtained.

(i) $E\hat{c}(\hat{\delta}) \geq \gamma \geq c(\hat{\delta})$.

(ii) $\hat{c}(\hat{\delta}) \to \gamma$ a.s. and in quadratic mean when the sample size in the training data increase to ∞.

(iii) When $m = 2$, the bias of $\hat{c}(\hat{\delta})$ for estimating γ is at worst of order $1/\sqrt{n}$, where n is the size of the training data.

(iv) When $m = 2$, and the distributions are multinomial, $P(c(\hat{\delta}) = \gamma) \to 1$ as $n \to \infty$.

(v) Smith's reallocation estimate for PCC using $\hat{\delta}$ is equal to $\hat{c}(\hat{\delta})$.

(vi) Suppose $\mathcal{X} = \{x_1, \ldots, x_k\}$, enclosed in a finite interval. Consider partitions of this interval into k disjoint subintervals with x_i in exactly one sub-interval. Let C_M be the collection of all m-partitions $(B_1, \ldots, B_m)$ such that B_i is a union of at most k sub-intervals containing x_i. The rule in C_M which maximizes the proportion of training data correctly classified is the same as $\hat{\delta}$. Moreover, $c(\hat{\delta}) \to \gamma$ a.s.

(vii) When $m = 2$, and the distributions are multinomials, he discussed on some shortcomings of Lachenbruch's estimate and suggested some other estimates and studied their performances numerically.

(b) Multivariate Bernoulli distributions.

The random variable X is a $p \times 1$ vector and each component of X takes values 0 or 1.

Bahadur (1961):

$m = 2$. This paper gives some approximations to the log likelihood-ratio, e.g., normal approximation and approximations using various truncations of Bahadur's series representation for the probability function . Some approximations to Kullback-Leibler symmetric information measure J are also obtained. These approximations are useful when J is small, p is large, and the interdependence among the components of X is not appreciable.

Solomon (1960, 1961):

$m = 2$. This is a numerical study of the effectiveness (PMC) and relative comparisons among rules based on the sum of the components, Fisher's LDF, LR statistic, and some truncated functions obtained from Bahadur's series representation for the probability functions.

Hills (1967):

$m = 2$. This is concerned with the problem of estimating log(LR) at a given point $X = x_0$. The following estimates are suggested.

(i) $(r_1/n_1) / (r_2/n_2)$, where r_i is the number of observations in a sample of size n_i from π_i which equal x_0.

(ii) 'Near neighbor' estimate of order 1,

$$\left(\frac{r_1 + r_1'}{n_1} \right) / \left(\frac{r_2 + r_2'}{n_2} \right),$$

where r_i' is the number of 'near neighbors' in a sample of size n_i from π_i whose x-value differ from x_0 in only one component.

(iii) 'Near neighbor' estimates of order > 1.

The distributions of these estimates are studied numerically. A step-wise method for selecting components using Kullback-Leibler information measure J is suggested.

Elashoff et al. (1967):

$m = 2$. Fisher's LDF, two functions based on a logistic model, and a function based on the assumption of mutual independence of the components are considered as possible classification statistics. The effectiveness of these statistics is studied numerically.

Martin and Bradley (1972):

The probability function of X in π_i is taken as

$$p_i(x) = f(x)[1 + h_s(a_i, x)],$$

where h_s is a linear function of the orthogonal polynomials on the sample space of X. This paper deals with the estimation of a_i and f subject to some constraints.

(c) Parametric non-normal continuous-type distributions.

Cooper (1962, 1963):

The distribution of X in π_i is taken as a known multivariate distribution of Pearson type II or type VII. The LR statistic is studied.

Bhattacharya and Das Gupta (1964):

$m = 2$. The distribution of X in π_i is taken as a member of the one-parameter exponential family. A class of admissible Bayes rules is obtained.

Cooper (1965):

The p.d.f. of X in π_i is taken as

$$p_i(x) = A_i|\Sigma_i|^{-\frac{1}{2}} f_i[(Q_i(x))^{\frac{1}{2}}],$$

where Q_i is a positive definite quadratic form and $f_i(u)$ decreases as u increases from 0. The LR statistic is studied.

Day and Kerridge (1967):

The p.d.f. of X in π_i is taken as

$$p_i(x) = d_i \exp[-\tfrac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)]f(x).$$

Two cases are considered, namely, (i) $f(x) \equiv 1$, (ii) $\Sigma = I$ and $f(x) = 1$ if every component of x is either 0 or 1 and $f(x) = 0$, otherwise. The posterior probability of the hypothesis $H_i: \pi = \pi_i$, given $X = x$, is expressed as $\exp(x'b + c)/[1 + \exp(x'b + c)]$. This paper mainly deals with the maximum likelihood estimates of b and c. For classification, it incorporates the idea of 'doubtful' decision.

Anderson (1972):

For the m-population, the posterior probability of $H_i: \pi = \pi_i$, given $X = x$, is taken as

$$p(H_i|x) = \exp(\alpha_{i0} + x'\alpha_i)p(H_m|x),$$

$$p(H_m|x) = 1/[1 + \sum_{i=1}^{m-1} \exp(\alpha_{i0} + x'\alpha_i)].$$

This paper deals with the estimation of α's by the maximum-likelihood method.

(d) Other cases.

Kendall (1966):

Some heuristic rules are suggested based on categorization of data.

Marshall and Olkin (1968):

For their formulation (see Section 3) of the problem, X is considered as a binomial random variable with the probability of success Y which is distributed as the uniform distribution on (0, 1). The form of a Bayes rule is obtained.

References (6)

Johnson, P. O. (1950). The quantification of qualitative data in discriminant analysis. J. Amer. Statist. Assoc. 45 65-76.

Matusita, K. (1956). Decision rule, based on the distance, for the classification problem. Ann. Inst. Statist. Math. 8 67-77.

Chernoff, H. (1956). A classification problem. Stanford University, Department of Statistics Technical Report No. 33.

Wesler, O. (1959). A classification problem involving multinomials. Ann. Math. Statist. 30 128-133.

Linhart, H. (1959). Techniques for discriminant analysis with discrete variables. Metrika 2 138-149 (MR21-6067).

Solomon, H. (1960). Classification procedures based on dichotomous response vectors. Contrib. Probability and Statistics (Hotelling vol.), 414-423.

Solomon, H. (1961). Classification procedures based on dichotomous response vectors. Stud. Item Anal. Pred. (ed. H. Solomon), Stanford Univ. Press, Stanford, California, 177-186.

Bahadur, R. R. (1961). On classification based on responses to N dichotomous items. Ibid. 169-176.

Raiffa, H. (1961). Ibid. (See Ref. 3).

Cochran, W. G. and Hopkins, C. E. (1961). Some classification problems with multivariate qualitative data. Biometrics 17 10-32.

Takakura, S. (1962). Some statistical methods of classification by the theory of quantification. Proc. Inst. Statist. Math. Tokyo 9 81-105 (MR27-3063).

Cooper, P. W. (1962). See Ref. 5.

Cooper, P. W. (1963). Statistical classification with quadratic forms. Biometrika 50 439-448.

Bhattacharya, P. K. and Das Gupta, S. (1964). See Ref. 5.

Cooper, P. W. (1965). See Ref. 5.

Hills, M. (1966). See Ref. 5.

Bunke, O. (1966). Nichparametrische Klassifikations verfahren fur qualitative und quantitative Beobachtungen. Wiss Z. Humboldt Univ. Berlin Math. Naturwiss. Reihe 15 15-18. (MR36-1031).

Kendall, M. G. (1966). Discrimination and classification. Proc. Internat. Symp. Multiv. Anal. (ed. P. R. Krishnaiah), Academic Press, New York, 165-185.

Hills, M. (1967). Discrimination and allocation with discrete data. Applied Statist. 16 237-250.

Elashoff, J. D., Elashoff, R. M., and Goldman, G. E. (1967). On the choice of variables in classification problems with dichotomous variables. Biometrika 54 668-670.

Day, N. E. and Kerridge, D. F. (1967). A general maximum like lihood discriminant. Biometrics 23 313-323.

Gilbert, E. (1968). On discrimination using qualitative variables. J. Amer. Statist. Assoc. 63 1399-1412.

Marshall, A. W. and Olkin, I. (1968). See Ref. 3.

Glick, N. (1969). See Ref. 4.

Martin, D. C. and Bradley, R. A. (1972). Probability models, esti mation and classification for multivariate dichotomous populations. Biometrics 28 203-222.

Anderson, J. A. (1972). Separate sample logistic discrimination. Biometrika 59 19-36.

7. Nonparametric or "Distribution-free" Methods.

The so-called nonparametric or distribution-free methods are used in statistical inference when one is concerned with a wide class of distributions which usually cannot be expressed as a parametric family with a finite number of parameters. When a statement regarding the probability of a certain statistical inference remains valid for every member in a given family of distributions, we call that a distribution-free inference with respect to that family; in particular, if the distribution of a statistic (used for inference) is the same for every member of a family of underlying distributions of the random variables involved, we say that the statistic is distribution-free with respect to that family. In the classification problem sometimes we face a similar situation when we devise rules for a broad class of underlying distributions whose structures cannot be expressed in simple parametric forms. However, unlike the problems of testing hypothesis or estimation, "a classification problem cannot be distribution-free" (Anderson, 1966) in the broad sense.

The available work in this area can be classified broadly into three main categories:

1) Consider a "good" rule (generally taken as a Bayes and/or an admissible minimax) assuming that the distributions are known. In this rule, replace the c.d.f.'s or the p.d.f.'s by their respective sample estimates. The rule thus obtained will be called a "plug-in"rule.

2) Use the statistics involved in devising some well-known tests for the nonparametric two-sample or k-sample problems.

3) Some ad-hoc methods which are typical for the classification problems, e.g.,"minimum distance" rule.

In the literature, the main emphasis is (a) to study the asymptotic behavior (e.g., consistency, efficiency in some sense) of the rules, (b) to

obtain some bounds for the PCC of a given rule, and (c) to study the small-sample performance.

Rules with Density Estimates.

There are several papers in the literature describing different methods for estimating a p.d.f. and the properties of different estimates. The following papers are mentioned in this connection; these references may be found in Van Ryzin (1966), Fu (1968,book), Patrick (1972, book), and Glick (1972).

Rosenblatt (1956, Ann. Math. Statist.)

Parzen (1962, Ann. Math. Statist.)

Cencov (1962, Soviet Math.)

Watson and Leadbetter (1963, Ann. Math. Statist.)

Aizerman, Braverman and Rozonoer (1964, Autom. Rem. Control)-Potential function method.

Nadarya (1965, Theory of Prob. and Appl.)

Loftsgarden and Quesenberry (1965, Ann. Math. Statist.)

Van Ryzin (1965, See Ref. 7)

Cacoullos (1966, Ann. Math. Statist.)

Murthy (1966, Ist. Internat. Symp. Multiv. Anal.)

Tsypkin (1966, Autom. Telemekhanika)-Stochastic approximation method.

Kashyap and Blaydon (1968, IEEE Trans. Inform. Theory)

Moore and Henrichon (1969, Ann. Math. Statist.)

As mentioned earlier, estimates of p.d.f.'s are used to obtain a plug-in rule from a given rule which involves density functions. Suppose δ^* is a Bayes rule with respect to a prior distribution ξ, assuming that the densities in the m populations are known. Let $R(\xi, \delta)$ be the Bayes risk of a rule δ and $\hat{\delta}^*$ be the plug-in rule obtained from δ^* by replacing the densities by

their respective estimates (based on training sample). Van Ryzin (1966) defined the notion of "Bayes-risk consistent" by the following:

$$P[R(\xi, \hat{\delta}^*) - R(\xi, \delta^*) \geq \epsilon] \to 0$$

as the sample sizes in the training sample tend to ∞. Van Ryzin also defined the Bayes risk consistency of order α_N by the following:

$$P[q_N\{R(\xi, \hat{\delta}^*) - R(\xi, \delta^*)\} \geq \epsilon\alpha_N] \to 0$$

as N = minimum of the sample sizes $\to \infty$ and q_N is any sequence $\to 0$ as $N \to \infty$. With respect to these notions, he studied some plug-in rules with different density estimates. For related results, see Van Ryzin (1965).

Glick (1969, 1972) obtained some properties of non-randomized plug-in rules assuming that the training data come from a mixed population (with unknown mixture ratios). Let $\gamma(\delta)$ be the PCC of a rule δ and δ^* be the rule which maximizes $\gamma(\delta)$ assuming that the class-densities and the mixture ratios are known. Let $\hat{\gamma}(\delta)$ be a plug-in estimate of $\gamma(\delta)$ by replacing the densities by their respective estimates. Glick's results are as follows:

i) If $\hat{f}_i \to f_i$ (density in π_i) a.s. ($i = 1, \ldots, m$) as the sample size in the training data increases to ∞, then

$$\hat{\gamma}(\delta) \to \gamma(\delta) \quad \text{a.s.},$$

uniformly in the class of all rules (not based on training data).

ii) If $\hat{f}_i \to f_i$ a.s. (in probability),

$$\gamma(\hat{\delta}^*) \to \gamma(\delta^*)$$

$$\hat{\gamma}(\hat{\delta}^*) \to \gamma(\delta^*)$$

a.s. (in probability).

iii) If the density estimates are pointwise unbiased, then

$$E[\hat{\gamma}(\hat{\delta}^*)] \geq \gamma(\delta^*) \geq \gamma(\hat{\delta}^*).$$

For other results, see the books by Fu (1968) and Patrick (1972).

Fix and Hodges (1951) also considered the density-plug-in rules (of which the nearest neighbor rules have drawn much attention) and studied the consistency of such rules.

Bunke (1966) considered the plug-in rule $\hat{\delta}$ obtained from a restricted (the prior probability measures are restricted to a given class) minimax rule δ by replacing the distributions involved by the respective empirical c.d.f.'s. He showed that asymptotically the rule $\hat{\delta}$ has the same Bayes-minimax property.

Nearest Neighbor (NN) Rules.

In 1951, Fix and Hodges proposed a classification rule for the two-population problem based on nonparametric estimates of the p.d.f.'s. Their method of estimating a density f can be described as follows: Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s with the common p.d.f. f which is continuous at x. Let $\{S_n\}$ be a sequence of sets in the sample space with corresponding volumes $\{V_n\}$, such that

i) $\lim_{n\to\infty} \sup_{y\in S_n} \|x-y\| = 0$,

ii) $\lim_{n\to\infty} nV_n = \infty$.

Let K_n be the number of observations that lie in S_n. Then

$$\hat{f}(x) \equiv \frac{k_n}{nV_n} \xrightarrow{p} f(x)$$

when $k_n \to \infty$, $n \to \infty$. Rosenblatt (1956) used this approach for

$$S_n = \{y: \|x-y\| \leq h_n\}, \lim h_n = 0.$$

Parzen (1962) replaced this set S_n by kernels $K_n(y, x)$. More generally,

$$\hat{f}(x) \equiv \frac{k_n}{nV_n} ,$$

where $V_n = \int K_n(x, y)dy$, $K_n = n\int K_n(x, y)dF_n(y)$, and F_m is the empirical c.d.f. based on $X_1,\ldots, X_n$. Watson and Leadbetter (1963) determined the best kernel which minimizes the integral square error for some specific f. Fix and Hodges (1951) also considered the sets S_n which depend on the sample $X_1,\ldots, X_n$; they suggested that S_n be defined as a "ball" with respect to some distance function d, centered at x, just large enough to contain k observations. For the m-population problem, one may also consider m different sequences of such sets. These estimates were studied by Loftsgarden and Quesenberry (1965).

The K-NN rule, as proposed by Fix and Hodges (1951) is described as follows. Let $\{X_{ij}; j = 1,\ldots, n_i\}$ be a random sample from the i^{th} population, $i = 1,\ldots, m$. Let X be the observation to be classified. Consider a distance function d and order all the value $d(X_{ij}, X)$, $j = 1,\ldots, n_i$; $i = 1,\ldots, m$. The K-NN rule assigns X to the population π_i, if $K_i/n_i = \max_j (K_j/n_j)$ where K_j is the number of observations from π_i in the K observations "nearest" to x; ties may be resolved in some manner. For $m = 2$, $n_1 = n_2 = n$, they showed that the PCC's of the K-NN rule (with d as the Euclidean distance) tend to the respective PCC's of the "likelihood maximum" rule when $n \to \infty$, $K \equiv K_n \to \infty$, $K_n/n \to 0$. Fix and Hodges (1953) obtained the exact and asymptotic expression for the PMC's of the NN rule when $p = 1$, $K = 1, 3$ and the parent distributions are normal with the same covariance matrix. For this normal case, they (numerically) compared the NN-rule with the ML rule for $p = 1, 2$; $k = 1, 3$.

Cover and Hart (1967) considered the mixed-population case and proposed a K-NN rule which assigns X to the population π_i, if $K_i = \max_j K_j$. They showed, under mild regularity conditions, that when the sample space is a separable metric space, and the distributions admit densities with respect to a measure, the limiting Bayes risk (0-1 loss function) of their 1-NN rule is bounded below by R^* and bounded above by $R^*(2-R^*m/(m-1))$, where R^* is the minimum Bayes risk (assuming that the distributions are known). Another result of Cover and Hart is as follows: Let $X, X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.'s in a separable metric space. Then X_n' = nearest neighbor to X among $X_1, \ldots, X_n$, tends to X with probability 1 as $n \to \infty$. In a later paper, Cover (1968) studied the rate of convergence of the Bayes risk of their 1-NN rule. In the above notation, let $\gamma(X, X_n')$ be the conditional Bayes risk of the 1-NN rule, given X and X_n', and let $\gamma^*(X)$ be the conditional Bayes risk, given X, under complete knowledge of the distributions. Peterson (1970) studied the different modes of convergence of

$$\gamma(X, X_n') - 2\gamma^*(X)[1-\gamma^*(X)]$$

under appropriate conditions. In a recent paper, Goldstein (1972) has studied some asymptotic properties of the K_n-NN rules and obtained a consistent upper bound for its PMC.

In 1966, Whitney and Dwyer considered the K-NN rule (of Cover and Hart) when the observations in the training sample are correctly identified with probability $\beta > 1/2$. Hellman (1970) modified the K-NN rule of Cover and Hart such that if at least K' of the K nearest neighbors to X comes from the same population, then X is assigned to that population; otherwise, decision is withheld. Specht (1966) noted that if the densities (p-variate) in the Bayes rule (mixed-population case) are replaced by the corresponding Parzen's estimate with

$$K_n(x, y) = \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right)$$

then the plug-in rule is the same as the 1-NN rule of Cover and Hart for σ sufficiently small.

In 1966, Patrick proposed another NN-rule in a more general framework. He considered different distance functions d_i such that

$$\lim_{\epsilon \to 0} \left[\max_{y} \{\|x-y\|: d_i(x, y) < \epsilon\}\right] = 0,$$

and the set $\{y: d_i(y, x) = \epsilon\}$ has zero volume for all $\epsilon > 0$ and all x. He suggested the following estimate of $f_i(x)$:

$$\hat{f}_i(x) = \frac{K_i(x)}{(n_i+1)V_i},$$

where $K_i(x)$ is a positive-integer and V_i is the volume of a d_i-neighborhood S_{in} of X depending on the training sample. Using these estimates he proposed the plug-in rule, obtained from the Bayes rule. For the special case, Patrick's NN rule assigns X to π_i if the K^{th} nearest neighbor to X in the sample from π_i is closest to X than that for a sample from any other population. An excellent account of these NN rules is given in Patrick's book (1972); see also the paper by Patrick and Fisher (1970). Pelto (1969) studied some estimates of the PMC of a NN-rule.

Rules Based on Distances Between Empirical c.d.f.'s.

For classification into two discrete distributions Matusita (1956) proposed the minimum distance rule based on Matusita-distance between the empirical c.d.f.'s and obtained some lower bounds for the PCC's. (See also Section 6). Das Gupta (1964) considered the minimum distance rule (with arbitrary distance) for the m-population problem and showed the consistency of such rules under appropriate conditions. He also obtained a lower bound for PCC of such rules and specialized this to the minimum Kolmogorov-distance rule.

Best-of-Class Rules.

The systematic development of this concept is due to Glick (1969). Suppose that the observation X to be classified comes from a mixture of m distributions. Consider a collection φ of ordered m-partitions of the sample space; for any such ordered partition $S = (S_1, \ldots, S_m)$, $X \in S_i$ leads to the decision that X comes from the i^{th} population π_i. Let $\gamma(S)$ be the PCC of the rule S. Define

$$\gamma(\varphi) = \sup_{S \in \varphi} \gamma(S) .$$

Let $X_1, \ldots, X_N$ be a supervised training sample. Then the "reallocation estimate" of $\gamma(S)$ is given by

$$\tilde{\gamma}(S) = \sum_{i=1}^{m} \frac{n_i}{N} \int_{S_i} d\hat{F}_i(x)$$

where n_i is the number of observations from π_i and $\hat{F}_i$ is the corresponding empirical c.d.f. Define

$$\tilde{\gamma}(\varphi) = \sup_{S \in \varphi} \tilde{\gamma}(S).$$

If a rule $\tilde{S} \in \varphi$ exists such that $\tilde{\gamma}(\varphi) = \tilde{\gamma}(\tilde{S})$ then $\tilde{S}$ is called a "best-of-class" rule in φ. The results obtained by Glick (1969) are stated below:

i) $E(\tilde{\gamma}(\varphi)) \geq \gamma(\varphi)$.

ii) $\sup_{S \in \varphi} |\tilde{\gamma}(S) - \gamma(S)| \to 0$ a.s. as $N \to \infty$.

iii) Let H_ν be the collection of all subsets of the sample space which are intersections of at most ν open half spaces. Let $\varphi(\nu_1, \nu_2)$ be the collection of all ordered m partitions $S = (S_1, \ldots, S_m)$ such that for each i, either S_i or its complement is a union of at most ν_2 sets, each of which

is a member of H_{ν_1} or the complement of a member. Let $\varphi \subset \varphi(\nu_1, \nu_2)$ be a collection of ordered m-partitions. Then, as $N \to \infty$

(a) $\tilde{\gamma}(\varphi) \to \gamma(\varphi)$ a.s.

(b) $|\tilde{\gamma}(\varphi) - \gamma(\tilde{S})| \to 0$ a.s.

(c) $\gamma(\tilde{S}) \to \gamma(\varphi)$ a.s.

It is to be noted that these results tacitly assume the existence of $\tilde{S}$. For $m = 2$, the collection of all hyperplane partitions coincide with $\varphi(1,1)$. The collection of all "interval" m-partitions, denoted by φ_I, is a subset of $\varphi(2, 2)$. When $\varphi = \varphi_I$, $\gamma(\varphi) \geq \gamma(\tilde{S})$.

Stoller (1954)

assumed $m = 2$ and the two distributions are such that an interval partition is the best one. Restricting to the class of all interval partitions (with known order) he proved the results (i), (iii)(a), (iii)(c) of Glick only "in probability" instead of "a.s." Hudimoto (1956) also considered the special case treated by Stoller and obtained an upper bound for the c.d.f. of $|\tilde{\gamma}(S) - \gamma(S)|$, where S is a rule with a given cut-off point ξ. Furthermore, he showed that the cut-off point $\hat{\xi}$ corresponding to the best-in-class rule $\tilde{S}$ is a consistent estimate of ξ. In a later paper, Hudimoto (1957) gave better bounds for the distribution of $\tilde{\gamma}(S)$ and obtained lower bounds for the c.d.f.'s of $\tilde{\gamma}(S) - \gamma(S)$, $\tilde{\gamma}(\varphi) - \gamma(\varphi)$, where φ is the class of all (known) ordered interval partitions and $m = 2$.

Rules Based on Tolerance Regions.

The idea of using tolerance regions for classification was first suggested by Anderson (1966), although it is implicit in the work of Fix and Hodges (1951). For the univariate case, Anderson suggested some variations of NN rules;

vector observations may be "ranked" (using them to define blocks) and then a univariate method can be applied. Other heuristic methods proposed by Anderson are as follows. Use the pooled training sample to construct "blocks." An observation X is classified into π_i if the block to which X belongs is defined by majority of observations from π_i. For the two-population problem, construct two sets of blocks separately based on the observations from π_1 and π_2. Let B_1 and B_2 be the blocks in the two sets which contain X. Consider the number of observations from π_2 in B_1 and the number of observations from π_1 in B_2 and classify X according to the larger number.

Quesenberry and Gessaman (1968) also suggested to use tolearance regions for the m-population classification problem with 2^m-1 decisions (instead of m decisions) described below:

$\delta_{i_1,\ldots,i_s}$: decide $P \in \{P_{i_1},\ldots, P_{i_s}\}$, $s = 1,\ldots, m-1$

δ_0 : reverse judgement

where $(i_1,\ldots, i_s)$ is a subset of $(1,2,\ldots, m)$. For each j, sample observations from π_j are used to construct a tolerance region A_j for P_j. They suggested a decision rule obtained by partitioning the sample space using the standard union-intersection method with the A_j's. The PMC's may be controlled by appropriately choosing the number of blocks used for A_j $(j = 1,\ldots, m)$. When the underlying distributions have some appropriate structure, the tolerance regions A_j can be so chosen that the resulting rule δ is consistent with the rule δ^* (i.e., $P_j(\delta = \delta^*) \to 1$, for each j) which minimizes the probabilities of reserving judgement subject to the size restrictions for the PMC's under complete knowledge of $P_1,\ldots, P_m$. However, in practice, the information concerning the distributions may not be sufficient enough so as to construct the above rule δ. Anderson and Benning (1970) partially

resolved this difficulty by using clustering techniques to get information on the likelihood-ratios. Patrick and Fisher (1970) used tolerance regions for estimating p.d.f.'s and plug-in rules. (See the discussion on NN rules.) Gessaman and Gessaman (1972) suggested some procedures based on statistically equivalent blocks and studied them by Monte Carlo methods.

Rules Based on Ranks--Analogy With Rank Tests.

The idea of using the statistics in the standard nonparametric rank-tests for devising classification rules was suggested by Das Gupta (1962, 1964). Das Gupta considered a rule which decides $P = P_i$ if $|W_i|$ is the smaller of $|W_1|$ and $|W_2|$, where W_i is the Wilcoxon statistic based on samples from π and π_i; he proved that this rule is consistent. Hudimoto (1964) modified this rule by taking W_i instead of $|W_i|$ when $F_1(x) \geq F_2(x)$ for all x; he derived a bound for the PCC of this rule and in a later paper (1965) studied it when ties may be present. Kinderman (1972) proposed a class of rules based on linear rank statistics as follows: Suppose n observations are available from each of three populations π_0, π_1, π_2. Define $N = 3n$,

$$T_{nj} = n^{-1} \sum_{i=1}^{N} E_{Ni} L_{ji} , \quad j = 0,1,2,$$

where E_{Ni} is a sequence of scores and L_{ji} is 1 if the i^{th} ordered observation in the pooled sample is from π_j, and 0 otherwise. Kinderman's rule classifies the observations from π_0 into π_1, iff $2T_{n0} - T_{n1} - T_{n2} > 0$; he assumed that the distribution in π_2 differs from that in π_1 by a positive shift in translation. He computed the relative asymptotic efficiency (in Pitman's sense) of this rule with the rule obtained by replacing the T_{nj} by the corresponding sample mean of the observations (from π_j) and specialized his results to "Wilcoxon's rank-sum" scores and "normal" scores. Govindarajulu and Gupta (1972) considered similar linear rank statistics for the m-population problem when the sample sizes may be different and obtained a rule based on them which asymptotically controls the average(with respect to a known prior) PCC.

For the two-population problem, a sequential rule based on Mann-Whitney statistics was proposed and studied by Woinsky and Kurz (1969). (See Fu's book (1968), for some other nonparametric sequential rules.)

An Empirical Bayes Approach.

Johns (1961) considered the two-category classification problem when I is considered as a random variable and the two categories are defined by a partition of the I-space. (See Section 2.) Following the empirical Bayes approach, he proposed a rule δ_N based on a training sample of size N and showed that the Bayes risk of δ_N tends to the minimum Bayes risk computed under complete knowledge of the distribution of (X, I). He treated the following three cases: (i) X is discrete-valued (supervised training sample); (ii) X is of continuous type, (supervised training sample); (iii) X is discrete-valued (post-supervised training sample). It may be noted that when I is treated as a classificatory variable and the loss function is 0-1, his rules reduce to NN rules.

Selection of Variables.

On the basis of random samples from two p-variate distributions, Patrick and Fisher (1969) devised a method for obtaining a q-dimensional ($q < p$) linear subspace of R^P such that the two induced q-variate marginal distributions are most "separated." Their method is based on nonparametric estimates (Murthy's extension of Parzen's estimate) of the p.d.f.'s and a 'separation' or distance criterion. For related work, see Patrick's book (1972) and Meisel's paper (1971). A nonparametric sequential method for including additional variates for classification is given in Smith and Yau (1972). For other methods, see Fu's book (1968), Wu (1970), Davisson et al. (1970).

Other Results.

Suppose that the c.d.f. of X is F_i in π_i ($i = 1,2$), where F_i has the mean μ_i and the covariance matrix Σ. Recall that the maximum likelihood rule classifies X into π_1, iff

$$X'\Sigma^{-1}(\mu_1 - \mu_2) > \tfrac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$$

when the distributions F_i are N_p. Using the well-known one-sided Chebyshev-inequality, Zhezhel (1968) showed that the maximum PMC of such a rule is $(1 + \Delta^2/4)^{-1}$, where $\Delta^2 = \|\mu_1 - \mu_2\|_\Sigma^2$, for all possible such F_i's.

Albert (1963) considered the classification problem where the supports of X are S_1 and S_2 in π_1 and π_2, respectively, where S_i's are unknown disjoint subsets of a Hilbert space such that the convex hulls of S_i's are at a positive distance apart. Samples are drawn from $S_0 \cup S_1$ sequentially and at the n^{th} stage a decision rule is given based on post-supervised training sample such that the PMC's tends to 0 as $n \to \infty$.

References (7).

Aoyama, H. (1950). Ann. Inst. Statist. Math. See Ref. 1.

Fix, E. and Hodges, J. L. (1951). Nonparametric discrimination: Consistency properties. U.S. Air Force School of Aviation Medicine. Report No. 4. Randolph Field, Texas.

Fix, E. and Hodges, J. L. (1953). Nonparametric discrimination: Small sample properties. Ibid. Report No. 11.

Stoller, D. C. (1954). Univariate two-population distribution-free discrimination. Jour. Amer. Statist. Assoc. 49 770-777.

Matusita, K. (1956). Ann. Inst. Statist. Math. See Ref. 6.

Hudimoto, H. (1956). On the distribution-free classification of an individual into one of two groups. Ann. Inst. Statist. Math. 8 105-112.

Hudimoto, H. (1957). A note on the probability of the correct classification when the distributions are not specified. Ann. Inst. Statist. Math. 9 31-36.

Johns, M. V. (1961). An empirical Bayes approach to nonparametric two-way classification. Studies in Item Analysis and Prediction. Ed. H. Solomon. Stanford University Press, Stanford, California.

Das Gupta, S. (1962). Univ. of N.C., Chapel Hill Mimeo No. 333. See Ref. 5.

Albert, A. (1963). A mathematical theory of pattern recognition. Ann. Math. Statist. 34 284-299.

Das Gupta, S. (1964). Nonparametric classification rules. Sankhyā A 26 25-30.

Hudimoto, H. (1964). On a distribution-free two-way classification. Ann. Inst. Statist. Math. 16 247-253.

Hudimoto, H. (1964). On the classification I. The case of two populations. Proc. Inst. Statist. Math.,Tokyo 11 31-38 (MR-29).

Hudimoto, H. (1965). On the classification II. Proc. Inst. Statist. Math., Tokyo 12 273-276 (MR-32).

Van Ryzin, J. (1965). Nonparametric Bayesian decision procedures for (pattern) classification with stochastic learning. Proc. IV Prague Conf. on Information Theory, Statistical Decision Functions, and Random Processes.

Van Ryzin, J. (1966). Bayes risk consistency of classification procedures using density estimation. Sankhyā A 26 25-30.

Bunke, O. (1966). See Ref. 6.

Anderson, T. W. (1966). Some nonparametric multivariate procedures based on statistically equivalent blocks. Proc. Ist. Internat. Symp. Multiv. Anal. Ed. P. R. Krishnaiah. Academic Press, New York, 5-27.

Whitney, A. W. and Dwyer, S. J., III (1966). Performance and implementation of the K-nearest neighbor decision rule with incorrectly identified training samples. Proc. IV Annual Allerton Conf. on Circuit Theory and System Theory. Champaign, Illinois.

Specht, D. F. (1966). Generation of polynomial discriminant functions for pattern recognition. Presented at IEEE Pattern Recognition Workshop, Puerto Rico.

Patrick, E. A. (1966). Distribution-free, minimum conditional risk learning systems. Purdue Univ. School of Elec. Engin. Tech. Rept. EE66-18. Lafayette, Indiana.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE Trans. Inform. Theory. IT-16 26-31.

Fu, K. S. (1968). Academic Press. See Ref. 1 (books).

Cover, T. M. (1968). Rates of convergence for nearest neighbor procedures. Proc. Hawaii Internat. Conf. on System Sciences. 413-415.

Quesenberry, C. P. and Gessaman, M. P. (1968). Nonparametric discrimination using tolerance regions. Ann. Math. Statist. 39 664-673.

Zhezhel, Yu N. (1968). The efficiency of a linear discriminant function for arbitrary distributions. Engineering Cybernetics 6 107-111.

Pelto, C. R. (1969). Adaptive nonparametric classification. Technometrics 11 775-792.

Glick, N. (1969). Stanf. Univ. Tech. Report. See Ref. 4.

Kurz, L. and Woinsky, M. M. (1969). Sequential nonparametric two-way classification with prescribed maximum asymptotic error. Ann. Math. Statist. 40 445-455.

Patrick, E. A. and Fischer, F. P. (1969). Nonparametric feature selection. IEEE Trans. Inform. Theory IT-15 577-584.

Patrick, E. A. and Fischer, F. P. (1970). Generalized K nearest neighbor decision rule. Jour. Information and Control. 16 128-152.

Peterson, D. W. (1970). Some convergence properties of a nearest neighbor rule. IEEE Trans. Inform. Theory IT-16 26-31.

Anderson, M. W. and Benning, R. D. (1970). A distribution-free discrimination procedure based on clustering. IEEE Trans. Inform. Theory IT-16 541-548.

Davisson, L. D., Feustel, E. A. and Modestino, J. W. (1970). The effects of dependence on nonparametric detection. IEEE Trans. Inform. Theory IT-16 32-41.

Wee, W. G. (1970). On feature selection in a class of distribution-free pattern classifiers. IEEE Trans. Inform. Theory IT-16 47-55.

Hellman, M. E. (1970). The nearest neighbor classification rule with a reject option. Presented at the IEEE Internat. Convention on Information Theory, Holland.

Meisel, W. S. (1971). On nonparametric feature selection. IEEE Trans. Inform. Theory. IT-17 105-106.

Glick, N. (1972). Jour. Amer. Statist. Assoc. See Ref. 4.

Kinderman, A. (1972). Univ. of Minn. Tech. Rep. See Ref. 4.

Patrick, E. A. (1972). Prentice Hall. See Ref. 1 (books).

Goldstein, M. (1972). K_n-nearest neighbor classification. IEEE Trans. Inform. Theory. IT-18 627-630.

Smith, S. E. and Yau, S. S. (1972). Linear sequential pattern classification. IEEE Trans. Inform. Theory. IT-18 673-678.

Govindarajulu, Z. and Gupta, A. K. (1972). Certain nonparametric classification rules: Univariate case. Michigan Univ. Statist. Dept. Tech. Report 17.

Gessaman, M. P. and Gessaman, P. H. (1972). A comparison of some multivariate discrimination procedures. Jour. Amer. Statist. Assoc. 67 468-472.

8. Miscellaneous References

(a) On distance functions.

Pearson, K.(1926). Biometrika (See Ref. 1).

Frechet, M. (1929). Sur la distance de deux variable aleatoires C.R. Acad. Sci., Paris 188, 368-370.

Mahalanobis, P.C. (1930) J. Asiatic Soc., Bengal (See Ref. 1).

Mahalanobis, P.C. (1936) Proc. Nat. Inst. Sci., India (See Ref. 1)

Hoel, P.G. (1944). On statistical coefficients of likeness. Univ. Calif. Publ. Math. (MR-6) 2(1) 1-8.

Bhattacharya, A. (1946). On a measure of divergence between two multinomial populations. Sankhya. 7 401-406 (MR-8).

Rao, C.R. (1947). Nature (See Ref. 1).

Ivanovic, B.V. (1954). Sur la discrimination des ensembles statistiques. Publ. Inst. Statist., Univ. of Paris 3 207-269 (MR-16).

Adhikari, B.P. and Joshi, D.D. (1956). Distance, discrimination et resume exhaustif. Publ. Inst. Statist., Univ. of Paris (MR-19). 5 57-74 (MR-19).

Frechet, M. (1957). Sur la distance de deux lois de probabilite. C.R. Acad. Sci., Paris 244 689-692 (MR-18).

Frechet, M. (1959). Les definitions de la Somme et du product Scalaires en terms de distance dans un space abstrait. (avec supplement) Cal. Math. Soc. (Golden Jubilee vol.) 1 151-157, 159-160.

Kullback, S. (1959). Information theory and Statistics. Wiley (Dover-1968).

Samuel, E. and Bachi, R. (1964). Measure of distances of distribution functions and some applications. Metron 23 83-121.

Ali, S.M. and Silvey, S.D. (1966). A general class of coefficients of divergence of one distribution from another. Jour. Roy. Statist. Soc., Series B. 28 134-142.

Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications. Ann. Inst. Statist. Math 19 181-192.

(b) Clustering

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proc. V. Berkely Symposium on Probability and Statistics. 2 , Univ. of Calif.

(c) Review

Hodges. J.L. (1950). (See Ref. 1).

Miller, R.G. (1962). Statistical Prediction by discrimination analysis. Amer. Meteor. Soc., Boston.

Nagy, G. (1967). State of the art in pattern recognition. Proc. IEEE, 56 836-860.