This paper reviews the procedure of evolving statistical classification rules.

Selection of variables, methods of classification, selection of a decision rule, and the problem of analyzing effectiveness of the technique are considered.

The procedure is demonstrated computationally by means of an example.

Statistical classification techniques

by C. F. Kossack

problems of discernment theory

One of the important areas of statistical analysis is discrimination among different populations. A suggested term for this analysis is *discernment theory*. Within this area there are essentially three major types of problems. So much confusion in terminology exists that it seems appropriate to first mention each type of problem before considering the particular problem of *statistical classification techniques*.

1. *The identification or discrimination problem.* If for a given group of distinct populations there is available a multi-variate random variable, the identification problem is that of evolving an index or numerical measure making use of the multi-variate random variable in such a way that this index, or measure, when evaluated for each of the populations, "best" identifies or discriminates among the several populations. The classical approach to this problem is that of using a linear discriminant function as first introduced by R. A. Fisher.[1]

2. *The statistical sorting or numerical taxonomy problem.* If one has reason to believe that his multi-variate sample came from more than one distinct population, the statistical sorting problem is that of evolving a rule, based only on the information contained in the mixed sample, for sorting the sample observations into the several (the exact number may itself be unknown) distinct populations. An approach to this problem considering each variable to be of a simple attribute (zero or one) type has been developed by T. T. Tanimato.[2]

3. *The classification or diagnosis problem.* If for a given group of distinct multi-variate populations one has available separate samples, the classification problem is that of evolving a decision rule that would enable one to assign a new or additional observation or individual into his proper population if all one knows about the individual is his multi-variate observational vector and that he came from one of the given populations. It is this latter problem that is the concern of the present paper.

terminology

For the convenience of the reader who may wish to review terminology, the following glossary (of terms used) is inserted:

Maximum likelihood estimates. If $f(X_1, X_2, \cdots, X_m; \theta)$ is the probability density for random sample of size n drawn from a population with unknown parameter θ then the maximum likelihood estimate of θ is the number $\hat{\theta}$ such that

$$f(X_1, X_2, \cdots, X_n; \hat{\theta}) \geq f(X_1, X_2, \cdots, X_n; \theta')$$

where θ' is any other possible value of θ.

Multi-variate normal probability density function.

$$f(X_1, X_2, \cdots, X_k) = (1/2\pi)^{k/2} \sqrt{|\sigma^{ij}|} \exp\left[-(1/2) \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma^{ij}(X_i - \mu_i)(X_j - \mu_j)\right]$$

where the inverse of the matrix $||\sigma^{ij}||$ is the matrix of variances and covariances.

Multinomial probability density function. The multinomial distribution is associated with repeated trials of an event which can have more than two outcomes. Its functional form is:

$$f(X_1, X_2, \cdots, X_k) = \frac{n!}{X_1!X_2! \cdots X_n!} p_1^{X_1} p_2^{X_2} \cdots p_k^{X_k}.$$

Measurement scales.

Nominal or *classificatory scale.* When numbers or other symbols are used simply to classify an object or characteristic. Example: the designation of postal zones.

Ordinal or *ranking scale.* If a "greater than" relationship holds for all pairs of classes of a nominal scale, we have an ordinal scale. Example: ranks in the military service.

Interval scale. An ordinal scale in which the differences ("distances") between any two numbers on the scale have comparative meaning. Example: measurement of temperature.

Ratio scale. An interval scale in which the zero point can not be arbitrarily chosen. Example: measurement of mass (as contrasted with temperature where assignment of zero to different temperatures on alternative scales, e.g., centigrade and Farenheit, is permissible).

Random variable. A variable whose occurrence is governed by a probability density function.

Categorical type variable. A variable whose measurement is in the nominal or classificatory scale.

Testing a statistical hypothesis. A statistical hypothesis is any statement relative to the nature of the probability function of a random variable, and any rule based upon observational data that determines whether or not one rejects the hypothesis is a test of the statistical hypothesis. Example: for the normal probability density function,

$$f(X) = 1/\sqrt{2\pi} \exp\left[-\frac{(X - \mu)^2}{2}\right] dX,$$

an hypothesis might be "The mean is less than 50," and a test of this hypothesis the rule, "If the mean of a sample from the population exceeds 65, reject the hypothesis."

Most powerful test of a statistical hypothesis. The power of a test of a statistical hypothesis is the probability of rejecting the hypothesis when it is false, while the level of significance of a test is the probability of rejecting the hypothesis when it is true. A most powerful test is one that has maximum power relative to all tests of equal or lower level of significance.

Likelihood. Let $X_1, X_2, \cdots, X_n$ be a sample of size n from a population with probability density function $f(X, \theta)$. The likelihood of the sample is then the product: $L = f(X_1, \theta)f(X_2, \theta) \cdots f(X_n, \theta)$.

Non-parametric methods. Techniques for estimating parameters and testing hypotheses which require no assumption about the form of the probability density function are called non-parametric methods.

Studentized form of a statistic. A statistic is said to be "studentized" when the parameter values in the originally derived form of the statistic are replaced by their maximum likelihood estimates from an available sample. Example: if the original form was $S = \sqrt{\sum [(X_i - \mu)^2/n]}$, then its studentized form would be $s = \sqrt{\sum [(X_i - \bar{X})^2/n]}$, where μ = parameter (the population mean) and $\bar{X} = \sum X_i/n$ (the sample mean).

statistical classification

The theory of statistical classification deals with the problem of assigning one or more individuals to one of several possible groups or populations on the basis of a set of characteristics observed on them. Thus, the problem of classification can be considered as a special case or application of multi-variate decision theory. The nature of the observed characteristics may vary from problem to problem. In some cases they may be all of a measured type while in another situation the variables may all be of the simple categorical type of attributes in which each observation can take on but one of a finite number of distinct values or states. Siegel has noted[3] that "measurements may, in general, be from four scales: the nominal, ordinal, interval, and ratio scales. In any given multi-variate classification problem, the measurements may be of a mixture involving some or all of these types of variables." It should be expected that numerous approaches have been advanced as to how one should go about evolving a classification decision rule. It is the purpose of this paper to examine the general problem of statistical classification and then to discuss some of the proposed solutions in some detail.

It should be recognized that since the area of interest has been designated as *statistical* classification, this means that the decision rule must be based upon observational data available from samples from the several populations rather than on known population characteristics. Thus we assume that we have a sample of individuals from each population and for each of these individuals we have available the same set of observations as are available for the individual requiring classification.

Consider for illustration a well-known classification problem, that of a prospective student applying for admission by submitting credentials such as his high school records and in addition being given a battery of admission tests. These data become the multi-variate set of observations available on each applicant.

The problem is to classify, in advance, the applicant into the population to which he belongs, where the alternatives are the population of those students who can successfully complete college training and the population of students who will not complete the college courses successfully. Available to the admissions office are the same data on former students, some known to have completed and the remaining known not to have completed college.

There are many classification problems in science and industry. For example, in biometric investigations one may want to assign a skull found in archaeological excavations to some dynastic period on the basis of anthropometric measurements. A taxonomist may want to classify a plant specimen into its proper species on the basis of measurements on its roots, stems, leaves, and flowers. Manufactured articles may be accepted or rejected on the basis of certain measurements made to determine whether or not they conform to specifications. Personnel may be assigned to duties on the basis of their scores in a battery of tests given to each employee. These and many more are essentially problems in classification and in a general sense are problems of statistical discrimination.

historical background

The first attempt to study the discrimination problem statistically was made by Karl Pearson in 1921. In a paper[4] by M. L. Tildesley, Pearson introduced the concept of a coefficient of racial likeness to serve as a measure of the "distance" between populations. The coefficient was defined in terms of sample means and variances. The coefficient assumed that all the observations were of a measured type having, in fact, the same variance and being uncorrelated. The coefficient was used to determine the probability that samples came from one and the same population and in this sense served as a test of divergence rather than as a classification rule.

In 1925, P. C. Mahalanobis of the Calcutta School of Statistics introduced the concept of a "measure" of divergence between two populations and in 1928, in a paper[5] presented to the Indian Science Congress, he proposed a generalized distance function, D^2. The original form of D^2 involved only population means, variances, and covariances but the "studentized" form of the statistic was considered at some length by R. C. Bose and S. N. Roy in a series of papers.[6]

In 1936, R. A. Fisher initiated a new approach to the problem of discrimination and classification with the introduction of linear discriminant functional analysis. This approach led to a new method of deriving test criteria suitable to multiple variate situations. The principle behind the choice of a "discriminant function" is merely to reduce multi-variate problems to univariate problems, a process that has been found extremely useful in multivariate analysis. The problem is reduced to that of a single variable by choosing a linear combination of the original variables and then constructing a statistic suitable for the univariate case. In principle, the discriminant function need not be linear but can be enlarged

to cover any class of functions. But in practice, a linear function of the original or transformed variables is nearly always chosen because it does not lead to complex distribution problems.

Another classification technique that is closely related to that of Mahalanobis' D^2 and Fisher's linear discriminant function is the one obtained by the late Abraham Wald. In a paper[7] appearing in 1944, Wald made an important contribution by introducing the statistical classification problem. He considered the specific problem of classifying a single p-variate observation into one of two p-variate normal populations, Π_1 and Π_2, being given that the observation belongs to Π_1 or to Π_2. The classification problem is reduced to a problem in testing the hypothesis H_1: that the observation belongs to Π_1, against the alternative hypothesis H_2: that the observation belongs to Π_2. The fundamental lemma due to Neyman and Pearson provides a classification statistic to classify the observation in Π_1 or Π_2 in a manner that is "best," where the "best" manner of classification corresponds to the most powerful test of H_1 against H_2. In the first instance the population parameters are assumed to be known, so that the classification statistic, depending on the parameters, is exactly known. In general the parameters are not known and it is therefore required that they be estimated from samples from the populations.

content of paper

The next section of the paper outlines a step-by-step procedure which may be followed in deriving a classification rule. This is followed by an illustrative example which details the application of the steps to a particular problem. Finally the distribution and multi-population problems are considered.

Derivation of a classification rule

Let us now look at the steps required to evolve a classification rule. Statistical classification rules, in general, depend either upon the concept of likelihood where one considers the ratios of the likelihoods that the observation to be classified came from the suspect populations, or they depend upon the value of some classification statistic whose form is assumed and is evaluated for the individual requiring classification. The samples that are available from each population are used to estimate the likelihood ratios or the constants in the classification statistic, depending on which approach is being used.

There are four major steps that must be accomplished if one is to evolve a classification rule, in brief: selection of the variables, selection of the classification technique, selection of the decision rule, and an analysis of effectiveness. These we now consider.

selection of variables

The selection of the variables to be used in making the classification. Here one encounters problems such as whether or not to include in his observational vector variables of different types, how reliable each available variable can be measured or determined, the discrimination power of the variable relative to the populations of interest, the inter-relationship of the variables, and the cost of

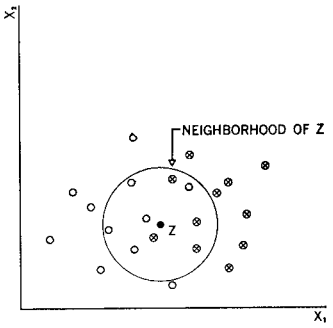
making each variable determination. The decisions of selection depend in the main on personal judgments since at present no good selection rule exists.

selection of classification technique

The technique to be used in making the classification estimate and the use of available sample data to make the estimates. One can identify several estimation techniques in the literature:

1. *Non-parametric estimation.* In this procedure the "closest" neighbors of the observation are identified in the pooled samples and the percentage of these that belong to each population is used as the likelihood ratio estimates. Figure 1 graphically displays the concept. Then the likelihood values for each population are estimated as $f_i(Z) = n_i/N$, where n_i = numbers of points in the neighborhood that are from the sample from population Π_i and N_i is the total number of observations available from Π_i.

Figure 1



2. *Classification by categories.* In this procedure, each variable is converted to a categorical type and one thus has states S_{1p}, S_{2p}, $\cdots$, $S_{n_p p}$ for the pth variable. All possible product classes are then identified, there being a total of $C = n_1 n_2 \cdots n_p$ different product classes. For example, suppose X_1 is the sex variable having the two states: S_{11} = male and S_{12} = female, and X_2 is the variable for age which has been converted to say, three states: S_{12} = young, S_{22} = middle age, S_{23} = old. Then there would be $2 \cdot 3 = 6$ product classes:

$C_1 = S_{11}S_{12}$ = young, males;
$C_2 = S_{21}S_{12}$ = young, females;
$C_3 = S_{11}S_{22}$ = middle aged, males;
$C_4 = S_{21}S_{22}$ = middle aged, females;
$C_5 = S_{11}S_{32}$ = old, males;
$C_6 = S_{21}S_{32}$ = old, females.

The frequencies $N_{\alpha i}$ are then determined, representing the number of individuals from the population Π_i who fall in the C_α class. The likelihood value for each class is then taken to be $f_i(C_\alpha) = n_{\alpha i}/N_i$.

3. *Parametric classification.* If a density function can be specified for each population, then the parameter vector $\theta = \theta_1, \theta_2, \cdots, \theta_m$ can be estimated for each population by using the maximum likelihood estimate obtained from each sample. Then the likelihood value associated with the observation Z for each population would be $L_i(Z) = f_i(Z; \hat{\theta}_i)$. Two of the most common density functions used are the multi-variate normal and the multinomial.

4. *Classification statistics.* The use of a classification statistic is usually restricted to the two-population case since one finds that most such statistics are formed through an algebraic simplification of the likelihood ratio function obtained by dividing the likelihood values for the two competitive populations. Such statistics usually involve the moments of the population distributions as well as the observational vector, Z. To evaluate the statistic

for a given Z, the maximum likelihood estimates of the parameters as obtained from the respective samples are used. Some of the better known statistics are:

a) The Wald Statistic:

$$W(Z) = \sum_{q=1}^{P} \sum_{p=1}^{P} \sigma^{pq}(\mu_q^{(2)} - \mu_q^{(1)})Z_p$$

where σ^{pq} = general term in the inverse of the common covariance matrix and $\mu_q^{(i)}$ = mean of X_q in population Π_i.

b) The Purdue Statistic:

$$B(Z) = \sum_{q=1}^{P} \sum_{p=1}^{P} [(\sigma_{(1)}^{pq} - \sigma_{(2)}^{pq})Z_pZ_q - 2(\sigma_{(1)}^{pq}\mu_p^{(1)} - \sigma_{(2)}^{pq}\mu_p^{(2)})Z_q + (\sigma_{(1)}^{pq}\mu_p^{(1)}\mu_q^{(1)} - \sigma_{(2)}^{pq}\mu_p^{(2)}\mu_q^{(2)})]$$

where index, i, indicates the characteristic for the ith population.

c) The Anderson Statistic:

$$A(Z) = a_1Z_1 + a_2Z_2 + \cdots + a_pZ_p + \cdots + a_PZ_P + b$$

where

$$a_p = \sum_{q=1}^{P} \sigma_{(r)}^{pq}(\mu_q^{(2)} - \mu_q^{(1)}),$$

$$b = \frac{\sqrt{a' \sum_{pq}^{(1)} a}\, a'\mu^{(1)} + \sqrt{a' \sum_{pq}^{(2)} a}\, a'\mu^{(2)}}{\sqrt{a' \sum_{pq}^{(2)} a} + \sqrt{a' \sum_{pq}^{(1)} a}},$$

and r is the positive root of the matrix equation

$$\{\textstyle\sum_r^{pq} \delta\}'\{\sum_{pq}^{(r^2)}\}\{\sum_{(r)}^{pq} \delta\} = 0,$$

where the $\sum$ matrices are

$$\textstyle\sum_{pq}^{(r)} = r \sum_{pq}^{(1)} + (1 - r) \sum_{pq}^{(2)}$$

and

$$\textstyle\sum_{pq}^{(r^2)} = r^2 \sum^{(1)} + (1 - r^2) \sum_{pq}^{(2)},$$

and the vector of mean differences, δ, is

$$\delta_p = \mu_p^{(2)} - \mu_p^{(1)}.$$

d) The Shaw Statistic

$$S(Z) = \sum_{q=1}^{P} \sum_{p=1}^{P} (\sigma_{(2)}^{pq} - \sigma_{(1)}^{pq})Z_pZ_q.$$

selection of decision rule

Selection of the decision rule to be used in making the actual classification decision for a given observation. To discuss this step at this stage it seems best to restrict our consideration to the two-population classification problem. We then have available for making the classification decision either a likelihood ratio that is a numerical function of the observational vector Z, say, $L(Z)$,

or we have a classification statistic defined as a numerical function of Z, say $C(Z)$. In either case a decision rule is then simply the division of the $L(Z)$ or $C(Z)$ one-dimensional interval into two regions such that for those Z's that yield an $L(Z)$ or $C(Z)$ that falls in region one the individual will be classified into population one; otherwise into population two. Thus we have reduced the problem of classification to that of determining the one region.

There are in general three decision strategies used in situations like this. The first relates to the control of the probabilities of making errors of misclassification. There are two such errors, the error of classifying an individual who really belongs to population Π_2 into population Π_1 and the error of classifying an individual who really relongs to Π_1 into population Π_2.

Since once we have designated one region of classification, the other region is automatically determined (being the complement of the first region with respect to the entire sample space), we can select the regions so as to control one of the errors of misclassification or so as to in some way balance the probability of one type error against the probability of making the other type error (i.e., making both probabilities equal), but we cannot exercise independent control on each error separately.

The second approach to selection of the classification region relates to the cost of making misclassification errors. We can determine the expected cost of misclassification through the formula:

$$C(R) = q_1 p(2 \mid 1, R)C(2 \mid 1) + q_2 p(1 \mid 2, R)C(1 \mid 2), \qquad \text{where}$$

q_i = the a priori probability of encountering an observation to be classified from Π_i,

$p(i \mid j, R)$ = the probability associated with the region R of classifying an individual into population Π_i given that he really belongs to Π_j, and

$C(i \mid j)$ = the cost of misclassifying an individual into population Π_i given that he really belongs to Π_j.

Usually one is interested in obtaining the region R that essentially minimizes this expected cost.

The third approach uses the so-called minimax approach found in decision making. In this case one seeks the region R that minimizes the maximum error that one may make. Often these requirements depend on numerical methods for their application. It seems best to defer a more detailed discussion of this step until an actual example is considered.

analysis of effectiveness

Determining the operational effectiveness of the classification technique. Basic to the measurement of the operational effectiveness of any classification technique are the probabilities:

$p(i \mid j)$ = the probability of misclassifying an individual who belongs in population Π_j into population Π_i.

From these probabilities one can evolve expected cost estimates as

well as other criteria of worth. To obtain estimates of these probabilities one requires the conditional distribution function of the likelihood ratios or the classification statistic used in the technique. In some cases these distributions can be expressed either exactly or approximately in mathematical form and then the misclassification probability estimations simply require the evaluation of an integral over the required region. When such a mathematical representation is not available, an empirical approach can be used involving the individual observations available in the samples to produce an empirical estimation of the conditional distributions. Here again it seems best to discuss the details of this step later around an actual problem.

Example

To demonstrate[8] how these approaches are actually utilized in a practical problem, let us consider the simple example of a classification problem, that of student admission to an engineering curriculum. Here we have the two populations—

Π_1: Students who would fail to do satisfactory work if admitted.
Π_2: Students who would do satisfactory work if admitted.

Let us consider for this example each of the four major steps required to evolve the classification rule.

step 1 *The selection of the variables to be used.* The problem of selection of variables in classification applications is comparable to that found in most scientific problems especially when the problem is being studied on an empirical or statistical basis. Thus one must not only select variables that form an adequate set for the discernment, but also must often consider techniques for reducing the original set of variables down to a more manageable set, since the use of a large number of variables in such problems frequently produces both arithmetic as well as theoretical complications. Included in this consideration is the possibility that transforms of the original variables may provide a better basis for making the classification than one would have if he used the variables in their original form. We will forego in this paper any further discussion of this fundamental problem and simply use for illustrative purposes the following three variables:

X_1 = The individual's score on a Mathematics Placement test.
X_2 = The individual's score on an English test.
X_3 = The individual's General Aptitude Test score.

step 2 *The technique to be used in making the classification.* Since the observational vector consists of three measured variables which we have reason to believe are distributed in each population as the multi-variate normal distribution with there being an equal covariance matrix for the two normal distributions, let us elect to use the Wald Classification Statistic as the technique. Thus

$$W(Z) = \sum_{q=1}^{3} \sum_{p=1}^{3} \sigma^{pq}(\mu_q^{(2)} - \mu_q^{(1)})Z_p,$$

and the classification rule that we will use will be:

"If $W(Z) > \lambda$ classify the Z as being in population Π_2."

(That is, we will admit the student to the curriculum.) Now $W(Z)$ is a linear function of Z which when the summation signs are expanded yields the form:

$$W(Z) = [\sigma^{11}(\mu_1^{(2)} - \mu^{(1)}_1) + \sigma^{12}(\mu_2^{(2)} - \mu_2^{(1)}) + \sigma^{13}(\mu_3^{(2)} - \mu_3^{(1)})]Z_1$$
$$+ [\sigma^{21}(\mu_1^{(2)} - \mu_1^{(1)}) + \sigma^{22}(\mu_2^{(2)} - \mu_2^{(1)}) + \sigma^{23}(\mu_3^{(2)} - \mu_3^{(1)})]Z$$
$$+ [\sigma^{31}(\mu_1^{(2)} - \mu_1^{(1)}) + \sigma^{32}(\mu_2^{(2)} - \mu_2^{(1)}) + \sigma^{33}(\mu_3^{(2)} - \mu_3^{(1)})]Z_3.$$

Thus to evaluate the coefficient of $W(Z)$ we will use as estimates of the σ_{pq}'s and the $\mu_p^{(1)}$ and $\mu_p^{(2)}$ the corresponding covariance and mean values obtained from the two samples. That is, the N_1 sets of triples (X_1, X_2, X_3) assumed available from unsuccessful students. Say that these samples have the numerical characters shown in Table 1. Then inverting the covariance matrix and sub-

Table 1

Sample means		Matrix of Pooled Covariances
Population 1	*Population 2*	
$\hat{\mu}_1^{(1)} = \bar{X}_1^{(1)} = 43.53$	$\hat{\mu}_1^{(2)} = X_1 = 60.26$	$\begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \hat{\sigma}_{23} \\ \hat{\sigma}_{31} & \hat{\sigma}_{32} & \hat{\mu}_{33} \end{bmatrix} = \begin{bmatrix} 154.39 & 63.15 & 35.03 \\ 63.15 & 110.93 & 27.49 \\ 35.03 & 27.49 & 39.19 \end{bmatrix}$
$\hat{\mu}_2^{(1)} = \bar{X}_2^{(1)} = 44.42$	$\hat{\mu}_2^{(2)} = X_2 = 59.50$	
$\hat{\mu}_3^{(1)} = \bar{X}_3^{(1)} = 16.37$	$\hat{\mu}_3^{(2)} = X_3 = 27.02$	

stituting these values into the equation for the Wald Statistic, we obtain

$$W(Z) = +0.0375Z_1 + 0.0672Z_2 + 0.1911Z_3.$$

Then if we were considering an individual for classification, say one whose scores on the three tests were

$$Z_1 = (34, 36, 12),$$

we would have the value of the statistic as

$$W(Z_1) = 0.0375(34) + 0.0672(36) + 0.1911(12) = 5.987.$$

The question of classification has then been reduced to that of deciding if $W(Z_1) = 5.987$ is greater than a prescribed λ.

step 3

Selection of the decision rule to be used in making the actual classification. In our student's admission problem we have seen how the problem was reduced to the question of how to determine the appropriate value of λ. Two of the three general decision approaches present themselves for consideration here: the *control of error approach* and the *cost control approach*. Let us consider each of these in turn.

1. *The control of error approach.* For our problem let us assume that the admissions office requires an admission policy such that the probability of a student doing unsuccessful work if admitted should be less than or equal to one tenth.

What is needed is the distribution function of the statistic $W(Z)$ since we would like to select λ such that $P\{W(Z) > \lambda \mid Z$ belongs to $\Pi_1\} = 0.10$. We know that $W(Z)$is asymptotically normally distributed under the condition that Z belongs to Π_1 with the mean,

$$\overline{W}_1 = \sum_{j=1}^{p} \sum_{i=1}^{p} \sigma^{ij}(\mu_j^{(2)} - \mu_j^{(1)})\mu_i^{(i)},$$

and variance,

$$V_W = \sum_{j=1}^{p} \sum_{i=1}^{p} \sigma^{ij}(\mu_i^{(2)} - \mu_i^{(1)})(\mu_j^{(2)} - \mu_j^{(1)}).$$

For the sample data whose characteristics were given above we find upon substituting the appropriate sample characteristics into the formula for the means and variance that

$\overline{W}_1 = 7.746$ and

$V_W = 3.676$.

Thus we have to solve for λ in the equation

$$p(2 \mid 1) = \frac{1}{\sqrt{2\pi}} \int_{(\lambda - \overline{W}_1)/\sqrt{V_W}}^{\infty} e^{-Z^2/2}\, dZ = 0.10.$$

From the table of areas under the normal curve we have

$$1.282 = \frac{\lambda - 7.746}{\sqrt{3.676}}$$

and

$\lambda = 10.20$,

and our classification decision rule can be stated as:

"If $W(Z) = +0.0350Z_1 + 0.0448Z_2 + 0.12747Z_3 > 10.20$ classify the observation as belonging to Π_2."

(That is, admit the student to the curriculum.)

In a more general sense we can balance the two values of the two misclassification probabilities by selecting the appropriate value of λ so as to meet any single constraint that might be imposed. For example, one may wish to control the errors such that two probabilities are equal. It is evident that the solution of the resulting integral equation may require a numerical technique of some sort.

2. *The cost control approach.* Consider in our student admission example that we have avilable the cost factors:

$C(2 \mid 1)$ = the cost of misclassifying an individual into population Π_2 when he really belongs to Π_1 (admitting a poor student) = 10, and

$C(1 \mid 2)$ = the cost of classifying an individual into population Π_1 when he belongs to Π_2 (failing to admit a good student) = 20.

Also assume we know:

q_1 = the a priori probability of a candidate for admission being from population $\Pi_1 = 0.25$,

q_2 = the a priori probability of a candidate for admission being from population $\Pi_2 = 0.75$.

Then if we wish an admission policy that would operate so as to minimize the expected loss, we have that

$$L_\lambda = q_1 p(2 \mid 1, \lambda) c(2 \mid 1) + q_2 p(1 \mid 2, \lambda) c(1 \mid 2)$$

where L_λ is the expected loss. In our particular case,

$$L_\lambda = (0.25)(10) p(2 \mid 1, \lambda) + (0.75)(20) p(1 \mid 2, \lambda)$$
$$= 2.5 p(2 \mid 1, \lambda) + 15.0 p(1 \mid 2, \lambda).$$

So we seek a λ which would minimize L_λ. One can simply try different values of λ, determine the $p(2 \mid 1, \lambda)$ and $p(1 \mid 2, \lambda)$ corresponding to the λ and then compute the L_λ. Since the relationship between L_λ and λ is quite smooth, one can through such a trial procedure approximate the appropriate minimizing value of λ within three or four steps.

step 4

Determining the effectiveness of the above classification rule. In the case of the above two populations—control of misclassification error situation—we compute the probabilities:

$p(2 \mid 1) = P$ {admitting a student who subsequently does unsatisfactory work}

$= P$ {classifying Z into Π_2 when Z belongs to Π_1},

and

$p(1 \mid 2) = P$ {failing to admit a student who could do successful work}

$= P$ {classifying Z into Π_1 when Z belongs to Π_2}.

Under Step 3 we determined the classification rule (i.e., the λ) such that $p(2 \mid 1) = 0.10$. To determine $p(1 \mid 2)$ we have

$$\overline{W}_2 = \sum_{j=1}^{P} \sum_{i=1}^{P} \sigma^{ij}(\mu_i^{(2)} - \mu_i^{(1)})\mu_i^{(2)} = 11.422,$$

and, due to the equal covariance assumption,

$$V_W = 3.676,$$

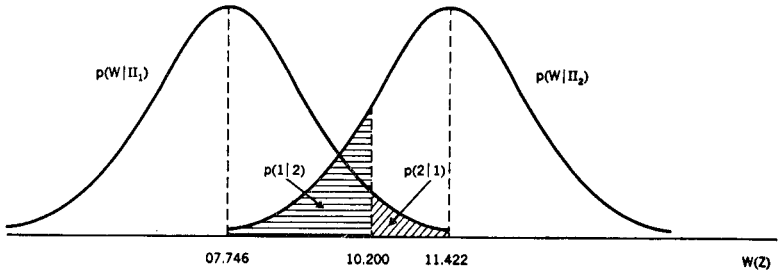
so

$$p(1 \mid 2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(10.20-11.422)/\sqrt{3.676}} e^{-Z^2/2}\, dZ = 0.26.$$

The rationale in these probability evaluations can best be exhibited graphically (Figure 2).

Thus we find that the operational effectiveness of the classification rule is such that $p(2 \mid 1) = 0.10$ and $p(1 \mid 2) = 0.26$. If

Figure 2 Probability evaluations



one is disturbed over the size of $p(1 \mid 2)$, he can either increase the allowable size of $p(2 \mid 1)$ or he may seek additional or new variables that better discriminate between the two populations.

Additional considerations

Essentially, each of the classification techniques identified above follow the four main developmental steps that were enumerated in detail for the Wald Classification Statistic. Two additional problems warrant special mention, however.

distribution problem

The first is the so-called distribution problem. That is, the requirement to have some knowledge as to how the statistic or likelihood ratio being used is distributed in probability under the condition that an individual comes from Π_1. This knowledge is required if one wants to formulate the particular classification rule to meet an error control or cost criterion. It is also needed if one is to estimate measures of operational effectiveness. We used the information that $W(Z)$ was normally distributed to generate these distribution requirements in the student admission illustrative example. One may, however, be interested in using a classification technique for which the mathematical form of its conditional probability distribution is unknown. In that case, especially if one has available a high speed digital computer and the sample sizes are sufficiently large, one can resort to the use of an empirically generated conditional distribution using the sample data. To illustrate the concept, let us suppose that we have available in the student admission problem data on 190 individuals known to be from population Π_1 (unsuccessful). Then if the value of the statistic, $W_1(Z)$, were computed for the 190 cases, these observations could be tabulated into a cumulated frequency distribution, the distribution plotted and a smooth distribution function drawn free-hand to approximate the ogive of the underlying conditional probability distribution. From such graphical representation appropriate values of $p(2 \mid 1, \lambda)$ and $p(1 \mid 2, \lambda)$ could be determined for corresponding values of λ. In our error control classification rule for the college admission problem we would have the frequency distribution and graphical representation as shown in Table 2 and Figure 3.

A comparable empirical estimate of the distribution of $W(Z)$ under the condition that the observation belongs to population Π_2 could be evolved through the use of the observation available in the sample from Π_2. The only variation in the technique would be in the accumulation of the frequencies. In this second case one would accumulate the frequencies with increasing W's.

Thus we would have an estimate of $p(2 \mid 1, \lambda)$ which yields the estimate of the probability of classifying an individual who is a Π_2 as a Π_1 if one used the decision rule "If $W(Z) > \lambda$ classify the individual into Π_2."

multi-population problem

The second problem that warrants additional mention is the multi-population problem. Here we are interested in classification procedures that could classify an individual into one of the several populations, where the number of populations is greater than two.

If one can associate with each population, Π_i, a q_i, the a priori probability of obtaining for classification an observation from population Π_i, and a cost factor, $C(j \mid i)$, associated with misclassifying an observation from Π_i as being from Π_j, then a decision rule is available that will minimize the expected cost of making classification. The rule states that:

"If

$$\sum_{i=1, \neq k}^{P} q_i p_i(Z) c(k \mid i) < \sum_{i=1, \neq j}^{P} q_i p_i(Z) c(j \mid i)$$

for all j $(j \neq k)$ then Z should be classified into Π_k."

If the inequality becomes an equality for some indices along with k, then it is immaterial as to whether the individual is classified into Π_k or one of the populations whose index yields the equality.

To illustrate the application of this rule, consider a three-population classification problem with

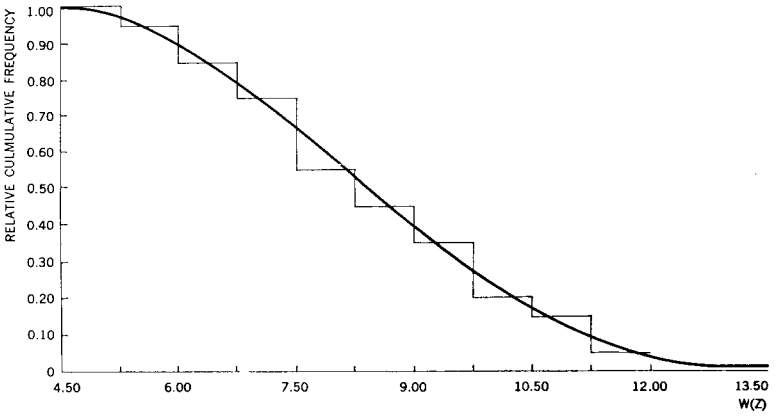
$$q_1 = 1/2, \quad q_2 = 1/3, \quad q_3 = 1/6$$

and the cost matrix $C(i\ j)$,

Table 2 Frequency distribution for $W_1(Z)$ college entrance problem, population Π_1 (Z)

Interval	*Tally*	*Cum.*	*%*
4.50– 5.24	9	190	100
5.25– 5.99	19	181	95
6.00– 6.74	24	162	85
6.75– 7.49	28	138	73
7.50– 8.24	21	110	58
8.25– 8.99	17	89	47
9.00– 9.74	24	72	38
9.75–10.49	12	38	20
10.50–11.24	18	24	13
11.25–11.99	2	6	03
12.00–12.74	4	4	02

Figure 3 Empirical distribution of W(Z) given Π_1



$$\begin{array}{c} \\ \Pi_1 \\ \Pi_2 \\ \Pi_3 \end{array} \begin{array}{c} \begin{array}{ccc} \Pi_1 & \Pi_2 & \Pi_3 \end{array} \\ \begin{bmatrix} - & 2 & 7 \\ 3 & - & 1 \\ 6 & 4 & - \end{bmatrix} \end{array}$$

with the Z to be classified having population likelihood values of $p_1(Z) = 0.40, p_2(Z) = 0.50, p_3(Z) = 0.25$.

Consider then the summations:

$i \neq 1, S_1 = q_2p_2(Z)C(1 \mid 2) + q_3p_3(Z)C(1 \mid 3);$

$i \neq 2, S_2 = q_1p_1(Z)C(2 \mid 1) + q_3p_3(Z)C(2 \mid 3);$

$i \neq 3, S_3 = q_1p_1(Z)C(3 \mid 1) + q_2p_2(Z)C(3 \mid 1).$

We have:

$S_1 = (\frac{1}{3})(0.50)(2) + (\frac{1}{6})(0.25)(7) = 0.61,$

$S_2 = (\frac{1}{2})(0.40)(3) + (\frac{1}{6})(0.25)(1) = 1.01,$

$S_3 = (\frac{1}{2})(0.40)(6) + (\frac{1}{3})(0.50)(6) = 2.20.$

And in this case, since S_1 is the smallest sum, we would classify the observation Z into Π_1.

It should be noted that if this method were to be utilized when no misclassification costs were available and one assumes that all the $C(j \mid i)$ are equal, say to unity, then the inequalities can be shown to reduce to:

"If

$q_kp_k(Z) < q_ip_i(Z) \quad \text{for all} \quad j \neq k,$

classify Z into Π_k, that is the most probable population."

Suppose that one does not have available a priori probabilities, then it is not possible to use the concept of minimum expected loss. In this case one of the decision strategies available is that of using a minimax solution, that is to obtain the classification decision rule that minimizes the maximum probability of making a misclassification error. To evolve this rule we may first consider the log likelihood functions:

$$U_{i,k}(Z) = \log \frac{p_i(Z)}{p_k(Z)}.$$

Now the inequalities, $U_{i,k}(Z) \geq C_i - C_k, k = 1, 2, \cdots, p, (k \neq j)$, with the C_k's being taken as non-negative will define a set of classification regions $R_1, R_2, \cdots, R_m$ in the sample space. To find the set of R's that yields the minimax solution, it is required that we use the C_r's such that the probabilities of correctly classifying an observation from Π_i into Π_i are equal for all i's. Here we have

$$p\{\Pi_i \mid \Pi_i, R\} = \int_{R_i} p_i(X)\, dX, \qquad i = 1, 2, \cdots, p.$$

The particular method of numerical evaluation of these integrals

for trial values of the C_r's would depend upon the assumed nature of the distribution functions $p_i(X)$.

concluding remarks

In conclusion it should be noted that the practical use of these classification techniques will usually require the use of high speed computing facilities. This is especially true if the dimension of the problem is at all large or if one must empirically generate the conditional distribution of the statistic being used by utilizing the individual observations available in the samples. There are many unresolved problems associated with the use of many of these techniques, but it is felt that the systematic exploration of their applicability in many practical problems cannot help but advance the general state of the art. Although the discriminating power of the set of variables currently being accumulated can be determined, the characteristics of the underlying distributions and the relative effectiveness of the competitive procedures must in many respects be tackled pragmatically. Attention must be given to the problem of estimating both the underlying a priori probabilities associated with the populations being considered along with the misclassification cost factors. Individuals may feel that such refinements are inappropriate to their particular classification problem, but it can be argued that until one addresses himself to the problem in some such systematic and scientific way, no real improvement can be expected. The criterion of worth of any system is its operational effectiveness and thus one should not only feel challenged to obtain estimates of the operational effectiveness of the "system" he is now using, but he should also investigate how the effectiveness may be improved by using one of the above statistical classification techniques.

CITED REFERENCES

1. Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," *Ann. Eugenics*, 7, pp. 179–188.
2. Rogers, D. J., and T. T. Tanimoto, "A Computer Program for Classifying Plants," *Science*, 132, pp. 1115–1123.
3. Siegel, S., "Non-parametric Statistics," Chapter 3, *McGraw-Hill*, 1956.
4. Tildesley, M. L., "A First Study of the Burmese Skull," *Biometrika*, vol. 13 (1921), pp. 176–262.
5. Mahalanobis, P. C., "On the Generalized Distance in Statistics," *Proc. Nat. Institute of Science* (India) vol. 12 (1936), pp. 49–55.
6. Bose, R. C. and S. N. Roy, "The Exact Distribution of the Studentized D^2-Statistic," *Sankhya* 4 (1938), pp. 19–38.
7. Wald, Abraham, "On the Statistical Problem Arising in the Classification of an Individual into One of Two Groups," *Annals of Math. Stat.*, vol. 15 (1944), pp. 145–162.
8. For details on the various approaches see: C. F. Kossack, "A Handbook on Classification," Purdue University Research Publication, 1962.

BIBLIOGRAPHY

- Anderson, T. W., "Classification by Multivariate Analysis," Psychometrika, 16, pp. 31–50.
- Anderson, T. W., "Introduction to Multivariate Statistical Analysis," Chapter 6, *John Wiley & Sons*, 1958.
- Cochran, W. G. and C. E. Hopkins, "Some Classification Problems with Multivariate Qualitative Data," *Biometrics*, Vol. 17, No. 1, March, 1961, pp. 10–32.

These papers introduce concepts involved in adapting the principal programming components within a single system.

After an examination of the over-all structure, the system's assembler, loader, and compilers are discussed. In this discussion (Parts I through V) attention is focused on the general design notions with minimal reference to the detail of mechanization and particular machines. Such reference, where necessary, is made to implementation of the system on the 7090.

Part VI compares implementation of the system on different machines and, to a certain extent, isolates the concepts that are independent of hardware.

Part VII is devoted to a general analysis of the system design.

Although some familiarity with the individual system components is assumed, an effort is made to address the systems engineer irrespective of his particular programming experience.

Design of an integrated programming and operating system

Part I: System considerations and the monitor

Part II: The assembly program and its language

Part III: Expanded function of the loader

Part IV: The system's FORTRAN compiler

Part V: The system's COBOL compiler

Part VI: Implementation on different machines

Part VII: Analysis of the system design

Parts I and II are published in this issue. The others will appear in successive issues.