SOME TECHNIQUES FOR SIMPLE CLASSIFICATION

CARL F. KOSSACK

UNIVERSITY OF OREGON

1. Introduction

In 1944 Wald[1] considered the problem of classifying a single multivariate observation, z, into one of two normally distributed parent populations, π_1 and π_2, when the only information available about the populations is contained in two samples of sizes N_1 and N_2, one drawn from each population. In order to obtain a classification technique, Wald assumed that the populations π_1 and π_2 have the same covariance matrix but unequal means and used the Neyman-Pearson[2] most powerful test for the hypothesis that z belongs to π_1 against the single alternative hypothesis that z belongs to π_2. The most powerful test for this hypothesis is given by the critical region $U \geqq d$, where $U = \sum_j \sum_i \sigma^{ij} z_i(\nu_j - \mu_j)$ and $\| \sigma^{ij} \|$ denotes the inverse matrix of the covariance matrix $\| \sigma_{ij} \|$, z_i the ith variate of the single observation, ν_j and μ_j the means of the jth variate for the populations π_1 and π_2. The critical region $U \geqq d$ is then approximated by $R \geqq d$, where R is the statistic obtained from U by replacing σ^{ij}, ν_j, and μ_j by their optimum estimates obtained from the two samples. In order to determine d corresponding to a given probability of an error of the first kind (classifying z in π_2 when z belongs to π_1) and the associated probabilty of an error of the second kind (classifying z in π_1 when z belongs to π_2) for the case when N_1 and N_2 are large, Wald used the fact that R can be approximated by means of the normal curve with means and covariance matrix obtained from the two samples.

In this paper we shall consider the problem of classifying an observation of a single variate into one of two normally distributed populations where the assumption of equal variances need not necessarily be valid. We shall distinguish this single-variate problem from the multivariate one by referring to it as simple classification.

2. Statement of the problem

We consider two variates x and y and assume that each is normally distributed and that each is independent of the other. A sample of size N_1 is drawn from the population π_1, the x-population, and a sample of size N_2 from the population π_2, the y-population. Denote by x_i the ith observation on x ($i = 1, 2, \cdots, N_1$) and by y_j the jth observation on y ($j = 1, 2, \cdots, N_2$). Denote by

[1] Abraham Wald, "On a statistical problem arising in the classification of an individual into one of two groups," *Annals of Math. Stat.*, vol. 15 (June, 1944).

[2] J. Neyman and E. S. Pearson, "Contributions to the theory of testing statistical hypotheses," *Stat. Res. Mem.*, vol. 1 (London, 1936).