ON A STATISTICAL PROBLEM ARISING IN THE CLASSIFICATION OF AN INDIVIDUAL INTO ONE OF TWO GROUPS[1]

By Abraham Wald

Columbia University

1. Introduction. In social, economic and industrial problems we are often confronted with the task of classifying an individual into one of two groups on the basis of a number of test scores. For example, in the case of personnel selection the acceptance or rejection of an applicant is frequently based on a number of test scores obtained by the applicant. A similar situation arises in connection with college entrance examinations. Again, on the basis of a number of test scores, the admission or rejection of a student has to be decided. In all such problems it is assumed that there are two populations, say π_1 and π_2, one representing the population of individuals fit, and the other the population of individuals unfit for the purpose under consideration. The problem is that of classifying an individual into one of the populations π_1 and π_2 on the basis of his test scores. Often, some statistical data from past experience are available which can be utilized in making the classification. Suppose that from past experience we have the test scores of N_1 individuals who *are known* to belong to population π_1, and also the test scores of N_2 individuals who *are known* to belong to population π_2. These data will be utilized in classifying a new individual on the basis of his test scores.

In this paper we shall deal with the statistical problem of classifying an individual into one of the populations π_1 and π_2 on the basis of his test scores and on the basis of past experience, given in the form of two samples, one drawn from π_1 and the other from π_2. In the next section we give a precise formulation of the statistical problem and state the assumptions we make about the populations π_1 and π_2.

2. Statement of the problem. We consider two sets of p variates $(x_1, \cdots, x_p)$ and $(y_1, \cdots, y_p)$. It is assumed that each of the sets $(x_1, \cdots, x_p)$ and $(y_1, \cdots, y_p)$ has a p-variate normal distribution and the two sets are independent of each other. It is furthermore assumed that the covariance matrix of the variates $x_1, \cdots, x_p$ is equal to the covariance matrix of the variates $y_1, \cdots, y_p$, i.e. $\sigma_{x_i x_j} = \sigma_{y_i y_j}$ $(i, j = 1, \cdots, p)$. We will denote this common covariance by σ_{ij}. Let us denote the mean value of x_i by μ_i and the mean value of y_i by ν_i. Furthermore we will denote the normal population with mean values $\mu_1, \cdots, \mu_p$ and covariance matrix $||\sigma_{ij}||$ by π_1, and the normal population with mean values $\nu_1, \cdots, \nu_p$ and covariance matrix $||\sigma_{ij}||$ by π_2.

A sample of size N_1 is drawn from the population π_1 and a sample of size N_2 is

---

[1] The author wishes to thank Dr. Irving Lorge, Columbia University, for calling his attention to this problem.