COMPARING DISTANCES BETWEEN MULTIVARIATE POPULATIONS—THE PROBLEM OF MINIMUM DISTANCE[1]

By M. S. Srivastava[2]

Princeton University

1. Introduction. For the problem of classification one assumes that the individual Π_0 to be classified belongs to one of the several given populations $\Pi_1, \Pi_2, \cdots, \Pi_k$. However, when the external evidence is slight, the classification problem is not only subject to the error due to the misclassification, but also to the error due to the false assumption that it (Π_0) belongs to one of the several given populations. The best thing would be, to first test whether Π_0 belongs to any one of the several given populations. If so, we assign Π_0 to the Π_i which corresponds to the hypothesis to be accepted at the highest level of significance. If we reject, we estimate the position of the new group relative to the others. Unfortunately, no such test criterion is available. Alternatively we might be interested to find which of the k population is 'closest' or 'nearest'—in the sense of distance, to the individual to be classified. This raises a natural question as to what measure of distance between two populations should be used. For multivariate populations, we shall use the Mahalanobis [3] generalized squared distance. Thus we are led to the investigation of the following problem. Given $k + 1$ populations $\Pi_0, \Pi_1, \cdots, \Pi_k$, to find which of the k populations $\Pi_1, \cdots, \Pi_k$ is nearest to Π_0. We consider in this paper, the case when Π_i's $i = 0, 1, \cdots, k$, are multivariate normal with means μ_i and common nonsingular covariance matrix Δ i.e. $\Pi_i : N(\mu_i, \Delta)$. The following example given by Cacoullos in [1] shows clearly the situation in which the above problem of nearest distance makes more sense than the classification approach.

Example. A p-dimensional observation X (e.g., the set of scores of a battery of p tests) is made on an individual; this individual is considered as a random observation from a certain category or population of individuals. A set of, say, k other populations is available. Each population may be thought of as a representative of a certain profession, and is characterized by a probability distribution of the p-measurements. The question is: which of the k populations does the individual fit best. If we introduce a measure of similarity between two professions, we are led to considering the problem of "nearest" (best fit) profession for the individual X.

The problem of nearest distance stems from Rao's paper [4], who suggested intuitively the maximum likelihood rule. When the mean μ_0 and the common covariance matrix Δ are both known, Cacoullos [1] proved the admissibility of

Received 26 August 1964; revised 20 September 1965.

[1] Work supported by the Office of Naval Research under Contract Non 1858(05) at Princeton University.

[2] On leave of absence from the University of Toronto.