A GENERAL THEORY OF DISCRIMINATION WHEN THE INFORMATION ABOUT ALTERNATIVE POPULATION DISTRIBUTIONS IS BASED ON SAMPLES

By C. Radhakrishna Rao

Indian Statistical Institute

1. Introduction. The problem of discrimination, that is of assigning an observed individual to its proper group, admits a simple solution when the distributions of measurements in the alternative populations are completely specified. Research in this direction originated with the use of the linear discriminant function introduced in 1936 by Fisher [3]. In 1939 Welch [24] showed that a general discriminant function in the case of two alternatives is the likelihood ratio of the two hypotheses, and is deducible either from Bayes' theorem with given a priori probabilities or by the use of a lemma by Neyman and Pearson [11] when the errors for the two hypotheses are minimised in any given ratio.

A general theory of decision functions when the alternatives are finite or infinite was developed by Wald [19] in 1939 and further generalized by him in 1949 [23]. In 1945 von Mises [9] obtained, in the case of a finite number of alternatives, the solution to the problem of minimising the maximum error, which is the general theme of Wald's work. Explicit solutions of Bayes' form, with given a priori probabilities or ratio of errors for the alternative groups, and the construction and use of a doubtful region were discussed by the author [13] in 1948. Related problems and the extension to problems of selection have been treated in a subsequent series of papers [15], [16].

In all these cases the alternative population distributions are assumed to be completely specified. The decision rule consists in setting up a correspondence between values observed in a sample and the alternative population distributions. In practice it is rarely possible to specify completely the distributions, but they may be estimable on the basis of independent samples from each of the alternative distributions.

Let $S_1, \cdots, S_k$ be independent samples from k alternative populations which may be partially specified, as when the functional forms of the probability densities are given but with unspecified parameters, or completely unspecified. After a sample S is drawn from a population known a priori to be one of the above set of k populations, the problem is to infer from which population the sample S has been drawn. The decision rule should be in the form of associating S with one of the samples $S_1, \cdots, S_k$, and declaring that S has come from the same population as the sample with which it is associated.

The usual practice is to estimate the alternative distributions on the basis of the sample information, and to use them in the solution which is strictly applicable when the alternatives are completely specified. This is probably the right

Received 8/6/53, revised 1/22/54.