SELECTION OF NONREPEATABLE OBSERVATIONS FOR ESTIMATION

G. ELFVING
UNIVERSITY OF HELSINGFORS AND COLUMBIA UNIVERSITY

1. Introduction and summary

This paper deals with the following particular type of design problem. Let there be given a set of possible observations, of the form

$$x_i = u_{i1}a_1 + \cdots + u_{ik}a_k + \xi_i = u_i'a + \xi_i\,, \qquad i = 1, \cdots, N\,, \tag{1.1}$$

where the coefficient vectors u_i are known, the parameter vector a is unknown, and the error terms ξ_i are uncorrelated random variables with mean 0 and variance 1. The last requirement can obviously be met by a change of scale if the original error variances are known. Let the aim of the investigator be to estimate a particular parametric form $\theta = c'a$. If it is required to do this on the basis of a subset comprising, say, $n < N$ observations, and if N and n are too large to permit trying out all possible combinations, one has to find some feasible selection procedure leading to a least-squares estimator $\hat{\theta}$ with as small variance as possible.

A practical situation where this problem is encountered is the following one, arising in psychology. Let the x_i's be the scores associated with various possible test items, and assume that a factor analysis has been performed, yielding a more or less approximate representation of the scores in terms of certain common factors $a_1, \cdots, a_k$ and mutually uncorrelated specific factors ξ_i. If the scores are normalized to specific variance one, and if the common factors are considered as parameters characteristic of the individual, we are concerned with the model (1.1). Further, let z be a "criterion score" measuring, for example, some ability of particular interest, for which, by the same factor analysis, a representation $z = c'a + \zeta$ has been found. For practical reasons, a planned routine prediction of z often has to be based on a moderate size subset of the original large set of items. The question then arises how to select this subset.

Our problem is closely connected with the allocation problem which, in its simplest form, can be stated as follows. Given a set (1.1) of possible observations, each of which can be independently *repeated* as many times as we please, which of them should we select for estimating $\theta = c'a$, and how many times should we repeat the selected ones when a fixed total n of actual observations is allowed? This problem may be considered as a special case of the previous one, namely, the case that all different coefficient vectors u_i occur in the given set with at least multiplicity n. This is approximately the situation when the "item points" u_i, in k-space, appear in clusters. In such a case it is possible to make use of certain geometric allocation methods developed by the writer [1], [2].

In the present paper, we shall be concerned with the opposite situation where the u_i's are more or less smoothly distributed, so as to permit an idealized description by means

This work was sponsored by the School of Aviation Medicine, Randolph Field, Texas, under Contract AF 18(600)-941.