ON THE EFFECT OF SELECTION PERFORMED ON SOME COORDINATES OF A MULTI-DIMENSIONAL POPULATION*

Z. W. BIRNBAUM, E. PAULSON, AND F. C. ANDREWS
LABORATORY OF STATISTICAL RESEARCH, UNIVERSITY OF WASHINGTON

Samples are often obtained under circumstances which make it less likely to draw individuals from some parts of a population than from others. In this paper a method is presented which makes it possible, under certain assumptions, to correct for the resulting bias and thus reconstruct the means, standard deviations, and correlation coefficients of the original population.

1. Introduction

In obtaining samples from a multi-dimensional universe, one often is limited to a part of the universe only, or finds it easier to obtain values from some part of the universe than values from the remaining part. If, for example, a number of psychological traits is being studied, it is often not possible to test individuals drawn at random from a universe, since one may have access only to individuals who have passed certain admission tests or have undergone a screening process. This situation occurs in schools, armed forces, institutional populations, etc. If it is desired to estimate the means, variances, and correlation coefficients of the universe, using such a sample, one clearly has to make corrections for the biased manner in which the sample was obtained.

It is the aim of this paper to present a general technique for solving problems of this kind. While the formal mathematical theorem referred to is not new, its application for solving the problem of estimating population parameters from biased samples seems not to be generally known† and deserves a detailed presentation.

This type of problem was known at least as early as 1901, at which time Karl Pearson (6) formulated it and gave a method for solving it in the special case in which the original universe was multi-normal and in which a selection was performed on some of the coordinates in such a manner that the resulting population was again

*Research done under the sponsorship of the Office of Naval Research.
†See Campbell (2) and the discussion in Cochran (3).

191
multi-normal. Pearson's method involved the laborious solution of systems of simultaneous linear equations. Taking up Pearson's problem, Aitken (1) set up the formal theory, again for the multinormal case, in the language of matrix algebra. A paper by Lawley followed (5), showing that the condition of normality can be relaxed. It is Lawley's theorem which will be used as the mathematical starting point of this presentation.

2. Notations. Formulation of the problem

Before the general problem of this paper can be formulated adequately, it appears necessary to introduce the following notations and definitions.

We consider the \( n \)-dimensional continuous random variable

\[
U = (X_1, X_2, \ldots, X_p, Y_{p+1}, Y_{p+2}, \ldots, Y_n)
\]

and denote by

\[
X = (X_1, X_2, \ldots, X_p)
\]

the \( p \)-dimensional random variable of those coordinates of \( U \) which will undergo selection, and by

\[
Y = (Y_{p+1}, Y_{p+2}, \ldots, Y_n)
\]

the \( (n-p) \)-dimensional random variable of the remaining coordinates of \( U \).

Let the joint probability density of all coordinates of \( U \) in the parent population \( \pi \) be

\[
f(U) = f(X, Y) = f(X_1, X_2, \ldots, X_p, Y_{p+1}, \ldots, Y_n),
\]

the marginal probability density of \( X \) in \( \pi \)

\[
g(X) = g(X_1, X_2, \ldots, X_p),
\]

and the conditional probability density in \( \pi \) of \( Y \) for a given \( X \)

\[
h(Y/X) = h(Y_{p+1}, Y_{p+2}, \ldots Y_n/X_1, X_2, \ldots, X_p),
\]

so that we have

\[
f(U) = f(X, Y) = g(X) \cdot h(Y/X).
\] (2.1)

The variance-covariance matrix of \( U \) in \( \pi \) will be denoted by

\[
V = ||\sigma_{ij}|| = \begin{bmatrix}
V_{p,p} & V_{p,n-p} \\
V_{n-p,p} & V_{n-p,n-p}
\end{bmatrix}
\] (2.2)