UNIVARIATE SELECTION: THE EFFECTS OF SIZE OF CORRELATION, DEGREE OF SKEW, AND DEGREE OF RESTRICTION

JAMES K. BREWER* AND JOHN R. HILLS

FLORIDA STATE UNIVERSITY

Pearson's formula for univariate selection was derived with the assumption of normality of variates before and after selection. This study examined the influence of skew upon estimates from Pearson's formula under certain conditions. It was found that even with essentially symmetric distributions, a large proportion of the data is necessary to obtain reasonably precise estimates of low correlations. With increasing skew, estimates become increasingly erroneous, the direction of the error depending upon which tail of the distribution is the basis of the estimates. Difficulties in applying correction for univariate selection in several studies of the predictability of college-grades for Negroes from scores on standard aptitude tests are discussed.

Karl Pearson [1902] provided procedures for estimating correlations for groups of different variability from a single population. He assumed that the variates were such as to be normally distributed both before and after selection. Lawley [1943] showed that Pearson's formulae were true under the more general conditions that the regression of y on x (where x is the predictor) is linear and that the variances and covariances of the distribution of y for a fixed set of values of x are constant. Rydberg [1963] reminds us that Pearson also recognized later that his formulae did not depend on frequencies being Gaussian in character.

No discussion has been found concerning the extent to which lack of symmetry may be ignored with relative impunity in using Pearson's procedures or what effects can be expected from violation of symmetry. In this study we evaluate the effect of various levels of positive skew of the predictor variable.

Such considerations are of practical importance in the context of several recent studies concerning the prediction of college performance from standardized aptitude tests in situations in which the aptitude tests are too difficult for the examinees [Biaggio, 1966; Hills & Gladney, 1968; Munday, 1965; Stanley & Porter, 1967]. This results in distributions that are seriously skewed positively.

*This manuscript was completed while the senior author was on an NSF Science Faculty Fellowship at Stanford University.
Procedure

In order empirically to investigate the influence of skew on corrections by Pearson's formula, two virtually symmetric distributions were generated by a procedure based on the Central Limit Theorem, and correlation was introduced between these distributions. One distribution was then made skew, and the corrected correlation was computed for each cumulative tenth of cases from each direction (i.e., left to right and right to left) of the predictor variable.

The most frequently encountered situation for adjustment for univariate selection was chosen as our frame of reference, i.e., explicit selection in the predictor variable. The formula for correcting for selection was

\[
R_{xy} = \frac{S_{xy}}{\sqrt{S_{x}^2 r_{xy}^2 + s_x^2 - s_{xy}^2}},
\]

(1)

where

- \( R_{xy} \) is the correlation in the extended (unselected) group,
- \( S_x \) is the standard deviation of \( X \) (the predictor variable on which selection has taken place) in the extended group,
- \( r_{xy} \) is the correlation in the curtailed group, and
- \( s_x \) is the standard deviation of \( x \) in the curtailed group.

The formula and notation were taken from Gulliksen [1950, p. 137, expression number 18]. This is the model used by Biagio, Hills and Gladney, and Munday.

Each of the symmetric distributions consisted of 1000 averages of 100 random two-digit numbers. The random numbers were produced by the random number generator in the CDC 6400 computer at Florida State University. This was similar to a procedure used by Boneau [1960] in investigating the robustness of the "t" distribution.

Various degrees of correlation were introduced in the following manner. Let \( S_i \), and \( S_{i+} \) \((i = 1, \ldots, 1000)\) represent the two symmetric distributions, and let \( Z_i \) \((i = 1, \ldots, 1000)\) represent a third symmetric distribution of two-digit numbers. Then the distributions given by

\[
y_i = \frac{aZ_i + bS_i}{a + b} \quad \text{and}
\]

(2)

\[
x_i = \frac{aZ_i + bS_i}{a + b}, \quad i = 1, 2, \ldots, 1000,
\]

(3)

will be correlated by an amount dependent on the values of \( a \) and \( b \). For example, if \( a = 3 \) and \( b = 1 \), then \( y_i \) and \( x_i \) will be correlated approximately .90. Using this procedure, initial correlations of approximately zero, .10, .20, .35, .50, .65, .80, and .90 were generated for the present investigation.