Florida Journal of Educational Research, 1973, 15

# THE EFFECT OF SPURIOUS HETEROGENIETY UPON CORRELATION

Wilson H. Guertin
and
William M. Fox

University of Florida

## SUMMARY

A shift in the calculated r from .72 to -.18 is illustrated for a problem with N = 50 when the error on a punched card is corrected.

## INTRODUCTION

When the scores of all but one subject in a sample lie within a narrow range on both variables being correlated, the pair of extreme scores will affect the r markedly. This follows since correlation is the ratio of covariance to total variance and that pair of scores is contributing most of the covariance as well as total variance. Guertin and Bailey warn about selecting "a homogeneous group with a few 'odd balls' added in"(1970, p. 173).

It now appears that these circumstances relating to variance are more likely to arise accidentally during computer analysis than because of an unfortunate selection of subjects. The rather remarkable effects of violating the rules are reported here as a warning to computer users.

Note the first pair of values in Table 1. They lie many standard deviations away from the means of each variable. They look as though they are spurious values and they are. They arose from a key punching error of improperly registering the card so that the punches fill in the wrong columns. We ran our data and found that the product moment r was an exciting .72. Dropping the bad card brought us sharply back to earth with the r going to the appropriate value of -.18.

At least three other circumstances can give rise to equally spurious computer output. Including a blank card accidentally within the data deck will cause the two scores to be read as 0.00's. The addition of such a blank card to the 44 other correctly punched cards causes the r = .18 to go to .73.

Such errors can be avoided by such practices as verifying the punched cards, making preliminary printed listings of the data cards to be run and marking the physically last card of the deck "LC" with a felt tipped marker.

It is very important to realize that erroneous zero values may be read from the data cards under two special circumstances. Missing data, indicated by blank fields on a card, can cause very serious distortion of results unless the program employed can ignore these missing observations instead of reading them as zeroes.

It should be remembered that an alphabetic card read in amongst the numeric data cards may not stop the run. Many systems will detect the inappropriate characters and convert them to numerical zeroes. The effect on the correlation computation is the same as if a blank card had been read in.

TABLE 1
Raw Scores as Punched on Cards and Corresponding Z-scores

| Raw Scores | | | | Standard Scores | | | |
|---|---|---|---|---|---|---|---|
| X | Y | X | Y | X | Y | X | Y |
| 0 | 3 | 37 | 82 | -5.64 | -5.83 | .02 | -.21 |
| 35 | 89 | 37 | 91 | -.29 | .29 | .02 | .44 |
| 35 | 91 | 37 | 87 | -.29 | .44 | .02 | .15 |
| 35 | 86 | 37 | 80 | -.29 | .08 | .02 | -.35 |
| 35 | 90 | 37 | 91 | -.29 | .36 | .02 | .44 |
| 35 | 98 | 37 | 92 | -.29 | .93 | .02 | .51 |
| 35 | 92 | 37 | 78 | -.29 | .51 | .02 | -.49 |
| 35 | 89 | 37 | 92 | -.29 | .29 | .02 | .51 |
| 35 | 95 | 37 | 78 | -.29 | .72 | .02 | -.49 |
| 36 | 77 | 38 | 93 | -.13 | -.56 | .17 | .58 |
| 36 | 78 | 38 | 83 | -.13 | -.49 | .17 | -.13 |
| 36 | 84 | 38 | 95 | -.13 | -.06 | .17 | .72 |
| 36 | 84 | 38 | 85 | -.13 | -.06 | .17 | .01 |
| 36 | 79 | 38 | 93 | -.13 | -.42 | .17 | .58 |
| 36 | 87 | 39 | 94 | -.13 | .15 | .33 | .65 |
| 36 | 95 | 40 | 78 | -.13 | .72 | .48 | -.49 |
| 36 | 81 | 40 | 75 | -.13 | -.28 | .48 | -.70 |
| 36 | 92 | 42 | 77 | -.13 | .51 | .78 | -.56 |
| 36 | 78 | 44 | 92 | -.13 | -.49 | 1.09 | .51 |
| 36 | 97 | 46 | 80 | -.13 | .86 | 1.40 | -.35 |
| 37 | 91 | 48 | 86 | .02 | .44 | 1.70 | .08 |
| 37 | 82 | 50 | 86 | .02 | -.21 | 2.01 | .08 |
| 37 | 94 | | | .02 | .65 | | |

## REFERENCES

Guertin, W. H. and Bailey, J. P. Introduction to modern factor analysis. Ann Arbor (Mich.): Edwards Bros., 1970.