EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1979, 39

A FORTRAN PROGRAM FOR TESTING AGREEMENT OF MULTIPLE OBSERVERS WITH A CATEGORICAL STANDARD ON NOMINAL SCALES

PAUL A. MCDERMOTT
University of Pennsylvania

MARLEY W. WATKINS
University of Nebraska-Lincoln

A computer program entitled Program STANDARD is presented which assesses the significance of the conjoint agreement of many observers with a standard or "correct" set of choices for assignments of subjects or objects on nominal scales.

IN psychological research and practice with data of a qualitative type, such as behavioral classifications, personality traits or diagnostic categories, it is usually considered important to assess the reliability with which observers assign such qualities, especially when the qualities are being attributed to human beings. To this end Cohen (1960) has developed the statistic κ, which may be used to test the significance of the agreement between *two* observers in their assignments of objects or subjects to nominal scales. Light (1971) and Fleiss (1971) later devised variations and extentions of κ, such as those represented by the statistic κ_m, that are appropriate for use in situations involving *multiple* (i.e., more than two) observers. In a recent article, Watkins and McDermott (1979) have described a computer program which applies Fleiss's (1971) statistical formulae for measuring levels of agreement among more than two observers in their assignments of subjects to categorical scales.

However, when considering the degree of interobserver consensus on nominal scales, it is often desirable to compare observers' choices with some *standard* or "correct" set of choices. This situation arises, for example, whenever the diagnostic decisions rendered by trainee clinicians must be compared with those of veteran clinicians or textbook ideals or when it is necessary to validate some categorical rating device by determining whether experts conjointly concur with the instrument's ratings. In other circumstances the nominal scale assignments of a particular observer among other observers may be of interest. In each of these situations a single set of categorical assignments is held as a criterion against which the categorical assignments of several raters must be evaluated. In such cases applications of κ_m would be inappropriate since κ_m reflects the *overall agreement of multiple observers with one another* and makes no assumption that any given standard set of choices might exist for comparisons with observers' assignments.

Guetzkow (1950) and Tukey (1950) proposed several approaches to testing the conjoint agreement of many observers with a standard. These methods were extended and refined by Light (1971) who defined the statistic G to test the null hypothesis of random assignments in multiple observers' categorizations relative to a "correct" choice for each assignment. Based upon a special variation of the contingency table routine, G compares the actual agreement of each of several observers with the standard category and then assesses this result in the light of what degree of agreement would have been expected by chance alone. For purposes of statistical testing, G is distributed approximately according to the unit normal variate.

Purpose

Within this context the purpose of this paper was to describe a computer program designated as Program STANDARD that tests the significance of the *conjoint agreement of multiple observers with a categorical standard on nominal scales* based upon Light's (1971) computational formulae for the statistic G.

Input

The job deck for each analysis consists of four control cards and an observer card deck which are arranged sequentially as follows:

Title card

An alphanumeric job title may be punched anywhere in columns 1–80.

Problem card

All numbers must be right adjusted.

Columns 1–3 = number of cases.
4–5 = number of categories.
6–8 = number of observers.

Standard cards

A standard or "correct" category choice must be indicated for each category specified in columns 4–5 of the problem card. Each category is designated by a category code where code 1 = 1st category, 2 = 2nd category . . . 9 = 9th category and 0 = 10th category.

Column 1 = code of standard category choice for 1st case.
2 = code of standard category choice for 2nd case.
. .
. .
. .
80 = code of standard category choice for 80th case.

Two standard cards must be provided. If the number of cases is greater than 80, one continues standard category specification on second standard card by punching code for 81st case in column 1, code for 82nd case in column 2, etc. If the number of cases is fewer than 81, the second standard card must be left blank.

Observer card deck

One or two observer cards must be punched for each observer.

Column 1 = code of category chosen for 1st case.
2 = code of category chosen for 2nd case.
. .
. .
. .
80 = code of category chosen for 80th case.

If the problem calls for more than 80 cases, two observer cards must be punched for each observer. One continues category choice specification by punching code for 81st case in column 1, code for 82nd case in column 2, etc. If the number of cases is fewer than 81, the second observer card for each observer must be omitted.

Output

The program provides the following information for each analysis:

1. *Job title* as specified by control card.
2. *Problem parameters* as specified by control and data cards.
   a. Number of cases, categories and observers.
   b. Standard category choices for each case.
   c. Observer category choices for each case.
3. *Value of G* and level of statistical significance if $p \leq .10$.

Capabilities and Limitations

Program STANDARD is written in FORTRAN IV for processing by computers in the IBM 360 series. It is fully documented with variables in mnenonic form corresponding to Light's (1971) computational formulae. Input editing and output specifications are provided for user's syntactical errors. At present, the program handles a maximum of 160 cases being assigned to as many as 10 categories by 100 or fewer observers. Any given observer may be used as the standard by substituting the observer's cards for standard cards.

Availability

A listing of the STANDARD source program, a copy of this paper, and a complete set of sample input and output data are available, without charge, from Dr. Paul A. McDermott, University of Pennsylvania, Graduate School of Education, 3700 Walnut Street, Philadelphia, Pennsylvania 19104.

REFERENCES

Cohen, J. A coefficient of agreement for nominal scales. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1960, 20, 37–46.

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 1971, 76, 378–382.

Guetzkow, H. Utilizing and categorizing problems in coding qualitative data. *Journal of Clinical Psychology*, 1950, 6, 47–58.

Light, R. J. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, 1971, 76, 365–377.

Tukey, J. W. Discussion on symposium. *Journal of Clinical Psychology*, 1950, 6, 61–74.

Watkins, M. W., and McDermott, P. A. A computer program for measuring levels of overall and partial congruence among multiple observers on nominal scales. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1979, 39, in press.