Behavior Research Methods & Instrumentation
1979, Vol. 11 (6), 607

A computer program for assessing conjoint interrater agreement with a correct set of classifications

PAUL A. McDERMOTT
University of Pennsylvania, Graduate School of Education, Philadelphia, Pennsylvania 19104

and

MARLEY W. WATKINS
Department of Educational Psychology and Measurements, University of Nebraska, Lincoln, Nebraska 68588

Several measures of response agreement for raters' classifications on nominal scales are finding increased popularity among researchers and practitioners. Such statistics (e.g., Cohen, 1960; Fleiss, 1971) share two principal features in common. First, they make no assumption that any true, correct, or standard set of nominal scale classifications might exist against which the classifications offered by the various raters could be evaluated. Second, in cases in which multiple-rater statistics are applied, the classifications of all raters are weighted equally, thus providing a measure of the overall conjoint agreement of the raters with each other and not with any existing correct or standard set of classifications.

When considering response agreement among raters in applied and research settings, it is frequently desirable to test the relative agreement among raters relative to a standard set of classifications. This is true, for example, whenever it is necessary to assess the categorizing ability of trainee observers in the light of an expert's categorizations or when the classification accuracy of a categorical rating device must be sized up against the conjoint ratings of independent expert observers. For this purpose, Light (1971) has defined the statistic G to test the significance of the conjoint agreement of many raters with a correct or standard set of classifications on nominal scales.

G is based upon a special version of a multiple contingency table routine that first compares the obtained agreement of each of the raters' categorical choices with the correct categorical choices and, thereafter, compares the observed level of agreement with what would be expected under the null hypothesis of random assignment of cases to categories. For purposes of testing statistical significance, G is distributed approximately according to the unit normal deviate.

The computer program described in this paper tests the statistical significance of the conjoint agreement of the categorical assignments of two or more raters with a correct set of classifications based upon Light's (1971) computational formulas for the statistic G.

Input. Each analysis requires four control cards and a data card deck as follows: (1) a title card; (2) a problem card to specify the number of cases being classified, number of categories, and number of raters; (3) a pair of standard cards indicating the correct category choice for each case considered; and (4) a set of observer cards, one or two for each rater, specifying raters' category choices for each case.

Output. The information provided for each analysis includes: (1) an alphanumeric job title; (2) number of cases, categories, and raters; (3) correct category choice for each case; (4) raters' category choices for each case; and (5) value of the G statistic and level of statistical significance associated with the unit normal deviate.

Computer and Language. The program is written in FORTRAN IV with ANSI standards for machines in the IBM 360/370 series and is adaptable to most other computer systems. Variables are in mnemonic form according to Light's (1971) computational formulas. Input editing and output specifications are provided for the user's syntactical errors.

Restrictions. Currently, the program will permit up to 160 cases to be assigned by 100 or fewer raters to a maximum of 10 categories.

Availability. A source listing, user's manual, and test input and output data may be obtained at no cost by writing to Paul A. McDermott, University of Pennsylvania, Graduate School of Education C1, 3700 Walnut Street, Philadelphia, Pennsylvania 19104.

REFERENCES

COHEN, J. A. 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.

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

(Accepted for publication July 26, 1979.)

0005-7878/79/060607-01$00.35/0