

MEASURING THE JOINT AGREEMENT BETWEEN
MULTIPLE RATERS AND A STANDARD

KENNETH J. BERRY AND PAUL W. MIELKE, JR.

Colorado State University

Reprint requests and correspondence should be addressed to Kenneth J. Berry, Department of Sociology, Colorado State University, Fort Collins, CO 80523-1784 or to Internet e-mail address: berry@lamar.colostate.edu.

MEASURING THE JOINT AGREEMENT BETWEEN MULTIPLE
RATERS AND A STANDARD

[Abstract]

A FORTRAN subroutine is presented to calculate a generalized measure of agreement between multiple raters and a set of correct responses at any level of measurement and among multiple responses, along with the associated probability value, under the null hypothesis.

MEASURING THE JOINT AGREEMENT BETWEEN MULTIPLE RATERS AND A STANDARD

A number of problems in educational and psychological research require the measurement of agreement between multiple raters and a standard or "correct set" of responses. Cohen (1960) introduced a chance-corrected index of agreement, kappa, which measured the agreement between two raters using categorical classification. Berry and Mielke (1988, 1990) extended Cohen's kappa to more than two raters, multivariate responses, and any level of measurement. Several investigators have considered measures of agreement for multiple raters and a correct set of responses. Guetzkow (1950) introduced procedures to measure agreement among multiple raters coding items when each item belongs to a known category. Tukey (1950) criticized this approach because of the assumption that each item is either coded correctly with 100 percent certainty, or it is coded at random. Light (1971) provided a test for the joint agreement of multiple raters with a correct set of responses and Hubert (1977) offered a "target rater" measure of agreement which is identical to the measure proposed by Light (1971), but with a much simpler formula for the variance.

The measures of agreement proposed by Guetzkow (1950), Light (1971), and Hubert (1977) are limited to categorical

data and a single response. In this paper, a chance-corrected index of agreement is presented which measures the agreement of multiple raters with a standard set of responses (which may be one of the raters). The proposed index can be used with any level of measurement and with multivariate responses. In addition, FORTRAN subroutine ASTAND is described which calculates the measure of agreement, \mathfrak{R} , and its associated probability.

Subroutine

If r is the number of responses/dimensions for each of n objects scored by m raters and the index of the standard set is denoted by s , then the measure of agreement, \mathfrak{R} , between the m raters and the standard set is defined by

$$\mathfrak{R} = 1 - \frac{\delta}{\mu_{\delta}} \quad (1)$$

with δ , the index of disagreement, given by

$$\delta = \sum_{i=1}^m \delta_i \quad (2)$$

where

$$\delta_i = \frac{1}{n} \sum_{j=1}^n \left[\sum_{k=1}^r (X_{sjk} - X_{ijk})^2 \right]^{\frac{1}{2}}, \quad (3)$$

for $i = 1, \dots, m$, and μ_δ is the subsequently defined expected value of δ under the null hypothesis, H_0 . If $m = 1$, $r = 1$, and the responses are categorical, then \mathfrak{K} reduces to the Cohen (1960) kappa statistic (Berry & Mielke, 1988).

Since \mathfrak{K} is a linear function of δ , the probability (P) of an observed \mathfrak{K} , \mathfrak{K}_o , under H_0 is the probability of an observed δ , δ_o , under H_0 given by

$$P(\mathfrak{K} \geq \mathfrak{K}_o | H_0) = P(\delta \leq \delta_o | H_0). \quad (4)$$

Under H_0 , each of the

$$M = (n!)^m \quad (5)$$

possible permutations of the m raters has an equal probability ($1/M$) of occurrence. Since the calculation of exact probability values is necessarily limited to small values of M , subroutine ASTAND computes approximate probability values based on the exact first three cumulants of the Pearson type III distribution. Because the m raters yield independent ratings, the exact mean, variance, and skewness of δ are given by

$$\mu_\delta = \sum_{i=1}^m \mu_i, \quad (6)$$

$$\sigma_{\delta}^2 = \sum_{i=1}^m \sigma_i^2, \quad (7)$$

and

$$\gamma_{\delta} = \frac{\sum_{i=1}^m c_3(\delta_i)}{\sigma_{\delta}^3}, \quad (8)$$

respectively, where μ_i , σ_i^2 , and $c_3(\delta_i)$ denote the first three exact cumulants of δ_i for $i = 1, \dots, m$. In order to use the Pearson type III distribution approximation, the standardized statistic

$$T = \frac{\delta - \mu_{\delta}}{\sigma_{\delta}} \quad (9)$$

accommodates the exact mean (μ_{δ}) and variance (σ_{δ}^2) and the Pearson type III distribution is characterized by the exact skewness (γ_{δ}) as given in Berry and Mielke (1988).

Subroutine Language

Subroutine ASTAND is written in ANSI FORTRAN-77 in double precision and runs on an IBM compatible PC. Comment lines provide input/output specification and documentation. Subroutine ASTAND is appropriate for analyses with $2 \leq m \leq 35$,

$2 \leq n \leq 50$, and $1 \leq r \leq 10$, but the dimensions may easily be changed by the user.

Input

The input into subroutine ASTAND consists of the values for the number of raters (m), the number of items to be rated (n), the number of responses/dimensions (r), and the data to be analyzed.

Output

The output includes the three input values, the values for δ , the expected value of δ (μ_δ), the variance of δ (σ_δ^2), the skewness of δ (γ_δ), the agreement measure (\mathfrak{R}), and the probability value (P).

Availability

A listing of subroutine ASTAND and an appropriate driver program is available from Kenneth J. Berry, Department of Sociology, Colorado State University, Fort Collins, CO 80523-1784, or by e-mail from Internet address: berry@lamar.colostate.edu.

REFERENCES

- Berry, K. J., & Mielke, P. W. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement, 48*, 921-933.
- Berry, K. J., & Mielke, P. W. (1990). A generalized agreement measure. *Educational and Psychological Measurement, 50*, 123-125.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.
- Guetzkow, H. (1950). Unitizing and categorizing problems in coding qualitative data. *Journal of Clinical Psychology, 6*, 47-58.
- Hubert, L. (1977). Kappa revisited. *Psychological Bulletin, 84*, 289-297.
- Light, R. J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin, 76*, 365-377.
- Tukey, J. W. (1950). Discussion on symposium on statistics for the clinician. *Journal of Clinical Psychology, 6*, 61-74.