

96/55.

AGREEMENT MEASURE COMPARISONS
BETWEEN TWO INDEPENDENT SETS
OF RATERS

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

Colorado State University

Address correspondence 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.

AGREEMENT MEASURE COMPARISONS BETWEEN TWO INDEPENDENT
SETS OF RATERS

[Abstract]

A FORTRAN program is described that calculates the probability of an observed difference between agreement measures obtained from two independent sets of raters.

AGREEMENT MEASURE COMPARISONS BETWEEN TWO INDEPENDENT
SETS OF RATERS

It is often of interest to evaluate the difference between measures of agreement obtained from two independent groups of raters. For example, if written essays are scored by a group of professional educators on a set of criteria (e.g., punctuation, grammar, etc.) and independently scored by a group of graduate students on the same set of criteria, it may be interesting to know (a) the amount of agreement among the professional educators, (b) the amount of agreement among the graduate students, and (c) the difference in agreement between the two groups. Berry and Mielke (1988, 1990) provide a generalized measure of agreement, \mathfrak{K} , for multiple raters, multivariate responses, and any level of measurement. The generalized agreement measure \mathfrak{K} may be used to evaluate the agreement among the professional educators and among the graduate students, but it cannot be used to evaluate the difference between the two sets of raters. In this paper, a test for difference between two independent \mathfrak{K} values is presented, and FORTRAN program DIFFER is described which calculates the difference, $D = \mathfrak{K}_1 - \mathfrak{K}_2$, and its associated probability.

Algorithm

Measure of Agreement

If c represents the number of responses for each of n objects scored by b raters, then the generalized measure of agreement, \mathfrak{A} , among the b raters is defined by

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

with δ , the index of disagreement, given by

$$\delta = \left[n \binom{b}{2} \right]^{-1} \sum_{i=1}^n \sum_{r < s} \Delta(\mathbf{x}_{ri}, \mathbf{x}_{si}), \quad (2)$$

where $\Delta(\mathbf{x}_{ri}, \mathbf{x}_{si})$ is a distance function given by

$$\Delta(\mathbf{x}_{ri}, \mathbf{x}_{si}) = \left[\sum_{k=1}^c (x_{rik} - x_{sik})^2 \right]^{1/2}, \quad (3)$$

b is the number of raters, x_{rik} denotes the k th element of vector \mathbf{x}_{ri} , with $r = 1, \dots, b$ and $i = 1, \dots, n$, and $\sum_{r < s}$ is the sum over all r and s such that $1 \leq r < s \leq b$ (Berry & Mielke, 1988, 1990).

Approximate probability values based on exact moments

of δ have been developed (Berry & Mielke, 1988). If δ_j denotes the j th value among $M = (n!)^b$ possible values of δ , the exact mean, variance, and skewness of δ , under the null hypothesis, are given by

$$\sigma_\delta^2 = M^{-1} \sum_{j=1}^M \delta_j^2 - \mu_\delta^2, \quad (4)$$

and

$$\gamma_\delta = \left(M^{-1} \sum_{j=1}^M \delta_j^3 - 3\mu_\delta \sigma_\delta^2 - \mu_\delta^3 \right) / \sigma_\delta^3, \quad (5)$$

respectively. Details of these measures are given in Berry and Mielke (1988) and in Mielke and Iyer (1982).

Difference Between Measures of Agreement

For convenience, let \mathfrak{R}_1 (\mathfrak{R}_2) denote the measure of agreement for group 1 (group 2) and let μ_1 (μ_2), σ_1^2 (σ_2^2), and γ_1 (γ_2) denote the mean, variance, and skewness for group 1 (group 2), respectively. Under the null hypothesis, H_0 , $\mathfrak{R}_1 = \mathfrak{R}_2$ and groups 1 and 2 are independent. If $D = \mathfrak{R}_1 - \mathfrak{R}_2$, then the exact mean, variance, and skewness of

D , under H_0 , are given by

$$\mu_D = 0, \quad (6)$$

$$\sigma_D^2 = \frac{\mu_1^2 \sigma_2^2 + \mu_2^2 \sigma_1^2}{\mu_1^2 \mu_2^2}, \quad (7)$$

and

$$\gamma_D = \frac{\mu_1^3 \sigma_2^3 \gamma_2 - \mu_2^3 \sigma_1^3 \gamma_1}{\mu_1^3 \mu_2^3 \sigma_D^3}, \quad (8)$$

respectively. The moment approximation incorporates the Pearson type III distribution which is characterized by the exact mean, variance, and skewness of D . The mean and variance of the Pearson type III distribution are 0 and 1, respectively. Since $\mu_D = 0$ under H_0 , the standardized test statistic, T , is given by

$$T = \frac{D}{\sigma_D}, \quad (9)$$

and $\gamma_T = \gamma_D$. If D_0 is the observed value of D , then the P-value associated with D_0 is the probability of having an

absolute value of D which is as large or larger than the absolute value of D_0 under H_0 .

It should be noted that although the computed moments are exact, the fact remains that the use of the Pearson type III distribution involves the continuous approximation of a discrete probability distribution. Consequently, a minimal sample size (n) of 10 is recommended to accommodate the approximation.

Program

Program DIFFER inputs values of \mathfrak{K}_1 , \mathfrak{K}_2 , μ_1 , μ_2 , σ_1^2 , σ_2^2 , γ_1 , and γ_2 , which are obtained from program AGREE (Berry & Mielke, 1990), and outputs values for D_0 , T , μ_D , σ_D^2 , γ_D , and the P-value.

Cicchetti and Heavens (1981) provide a Z statistic and Fortran program which evaluates the difference between two independent values of Cohen's (1960) kappa agreement measure and Cohen's (1968) weighted kappa agreement measure. While the \mathfrak{K} measure of agreement can be shown to reduce to Cohen's (1960) kappa measure of agreement when there are only two raters, a single subject response, and nominal level of measurement, \mathfrak{K} is a much more general measure of agreement which allows for multiple raters, multiple subject responses, and any level of measurement (Berry & Mielke, 1988). In addition, because the Z statistic of Cicchetti

and Heavens (1981) utilizes the standard error of kappa derived by Fleiss, Cohen, and Everitt (1969) and Fleiss and Cicchetti (1978), its inferential asymptotic solution requires large sample sizes before normality is achieved; otherwise, the Z statistic overestimates the true probability value. Cicchetti and Heavens (1981) recommend a minimum of $3k^2$ subjects, where k is the number of categories. Thus, for example, $k = 7$ categories would require a minimum of 147 subjects. In contrast, the probability of differences between two \mathfrak{K} values is based on an inferential nonasymptotic solution, and a sample size of only 10 is sufficient for all analyses of differences between two \mathfrak{K} measures in program DIFFER.

Example

Consider a set of $n = 40$ undergraduate essays evaluated by two independent sets of graders on six canons of writing excellence: clarity, organization, accuracy, spelling, grammar, and punctuation. One set of graders is composed of $b = 3$ faculty in English composition, and the second set of graders consists of $b = 8$ advanced graduate students in English composition. Each of the six attributes is scored on a scale of 1 to 10 by each of the graders. For the $b = 3$ faculty, $\mathfrak{K} = 0.1158$, $\mu_{\delta} = 1.2705$, $\sigma_{\delta}^2 = 0.4678 \times 10^{-3}$, and $\gamma_{\delta} = -0.3415$. For the $b = 8$ graduate students, $\mathfrak{K} = 0.1978$,

$\mu_{\delta} = 1.6024$, $\sigma_{\delta}^2 = 0.1010 \times 10^{-2}$, and $\gamma_{\delta} = -0.2843$. Analysis with program DIFFER yields $D = -0.0820$, $\sigma_D^2 = 0.6832 \times 10^{-3}$, $T = -3.1380$, $\gamma_D = -0.2985 \times 10^{-1}$, and the probability of a D this large or larger under $H_0: \mu_D = 0$ is 0.1966×10^{-2} .

Program Language

Program DIFFER is written in ANSI FORTRAN 77 in double precision and runs on an IBM compatible PC. Comment lines provide input/output specification and documentation.

Availability

Programs AGREE and DIFFER are available 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.
- Cicchetti, D. V., & Heavens, R. (1981). A computer program for determining the significance of the difference between pairs of independently derived values of kappa or weighted kappa. *Educational and Psychological Measurement, 41*, 189-193.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.
- Fleiss, J. L., & Cicchetti, D. V. (1978). Inference about weighted kappa in the non-null case. *Applied Psychological Measurement, 2*, 113-117.
- Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323-327.

Mielke, P. W., & Iyer, H. K. (1982). Permutation techniques for analyzing multiresponse data from randomized block experiments. *Communications in Statistics: Theory and Methods*, 11, 1427-1437.