PSYCHOMETRIKA—VOL. 43, NO. 2.
JUNE, 1978

APPROXIMATE INTERVAL ESTIMATION FOR A CERTAIN INTRACLASS CORRELATION COEFFICIENT

JOSEPH L. FLEISS AND PATRICK E. SHROUT

COLUMBIA UNIVERSITY

When the raters participating in a reliability study are a random sample from a larger population of raters, inferences about the intraclass correlation coefficient must be based on the three mean squares from the analysis of variance table summarizing the results: between subjects, between raters, and error. An approximate confidence interval for the parameter is presented as a function of these three mean squares.

Key words: confidence intervals, analysis of variance—random model, inter-rater reliability.

Suppose that a reliability study is conducted in which each of a random sample of I raters independently rates each of a random sample of N subjects. The rating by the ith rater on the nth subject, say X_{in}, is assumed to be representable as

$$X_{in} = \mu + r_i + s_n + e_{in}(i = 1, \cdots, I; n = 1, \cdots, N),$$

where μ is the overall mean level of rating; r_i is the effect of the ith rater (assumed to be normally distributed with mean 0 and variance σ_r^2); s_n is the effect of the nth subject (assumed to be normally distributed with mean 0 and variance σ_s^2); and e_{in} is the error associated with this particular rating (assumed to be normally distributed with mean 0 and variance σ_e^2). All random variables $\{r_i, s_n, e_{in}: i = 1, \cdots, I; n = 1, \cdots, N\}$ are assumed to be mutually independent. It is specifically assumed that there is no rater-by-subject interaction.

The variance of any single rating is equal to $\sigma_r^2 + \sigma_s^2 + \sigma_e^2$, and the covariance between the ratings by two randomly selected raters on a random subject is σ_s^2. The intraclass correlation coefficient of reliability is then equal to

$$\rho = \frac{\sigma_s^2}{\sigma_r^2 + \sigma_s^2 + \sigma_e^2}. \tag{1}$$

Let Table 1 represent the results of an analysis of variance applied to the IN ratings. It is easily checked from the column of expected mean squares that

$$\hat{\rho} = \frac{N(BMS - EMS)}{I \cdot RMS + N \cdot BMS + (IN - I - N)EMS} \tag{2}$$

is a ratio of unbiased estimates of the numerator and denominator of ρ; $\hat{\rho}$ is a consistent estimate of ρ as both I and N increase. The estimate (2) of ρ was, to our knowledge, first proposed by Bartko [1966].

The hypothesis that $\rho = 0$ is equivalent to the hypothesis that $\sigma_s^2 = 0$. This hypothesis is rejected if the ratio BMS/EMS exceeds the critical value of the F distribution with $(N - 1)$ and $(I - 1)(N - 1)$ degrees of freedom. A confidence interval for ρ, however, must be a

Dr. Fleiss is also with the Biometrics Research Unit of the New York State Psychiatric Institute. This work was supported in part by grant DE 04068 from the National Institute of Dental Research.

Reprint requests should be sent to Joseph L. Fleiss, Division of Biostatistics, Columbia University School of Public Health, 600 West 168 Street, New York, New York 10032.

TABLE 1

Analysis of Variance Table for a Random Effects Inter-rater Reliability Study

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Expected Mean Square
Between Subjects	N-1	$I \sum_{n=1}^{N}(\bar{X}_{.n}-\bar{X}_{..})^2$	BMS	$\sigma_e^2 + I\sigma_s^2$
Between Raters	I-1	$N \sum_{i=1}^{I}(\bar{X}_{i.}-\bar{X}_{..})^2$	RMS	$\sigma_e^2 + N\sigma_r^2$
Error	(N-1)(I-1)	$\sum_{i=1}^{I}\sum_{n=1}^{N}(X_{in}-\bar{X}_{i.}-\bar{X}_{.n}+\bar{X}_{..})^2$	EMS	σ_e^2

function not only of *BMS* and *EMS*, but of *RMS* as well. The construction of such an interval is a problem which apparently has not yet been addressed.

An approximate solution consists of applying the same kind of reasoning used, for example, by Feldt [1965] for the case where the estimated intraclass correlation coefficient involves only two independent mean squares. If ρ were known but the individual components of variance not, the expectation of the mean square between subjects could be expressed as

$$\mathcal{E}(BMS) = \frac{1}{1-\rho}\,(I\rho\,\sigma_r^2 + \{1 + (I-1)\rho\}\sigma_e^2)\,. \tag{3}$$

It is easily checked that the expectation of

$$V = \frac{1}{N(1-\rho)}\,(I\rho RMS + \{N(1 + (I-1)\rho) - I\rho\}EMS)$$

is also equal to the expression given in (3). The quantity V is distributed independently of *BMS*, but not exactly as a constant times a chi square variable.

Let $F_r = RMS/EMS$. Following Satterthwaite [1946], V can be shown to be approximately distributed as $c\chi_\nu^2/\nu$, where χ_ν^2 denotes a variable distributed as chi square with ν degrees of freedom, $c = \mathcal{E}(BMS)$, and

$$\nu = \frac{(I-1)(N-1)(I\rho F_r + N(1 + (I-1)\rho) - I\rho)^2}{(N-1)I^2\rho^2F_r^2 + (N(1 + (I-1)\rho) - I\rho)^2}\,. \tag{4}$$

Thus, the random variable BMS/V has, approximately, an F distribution with $(N-1)$ and ν degrees of freedom. Let ν be estimated from (4) with $\hat{\rho}$, defined in (2), replacing ρ. Then the approximate probability statement

$$1 - \alpha \doteq \Pr\left\{\frac{BMS}{V} < F^*\right\},$$

where F^* is the upper 100 $(1-\alpha)$ percentile of the F distribution with $(N-1)$ and ν degrees of freedom, may be converted, by simple algebra, into the approximate $100(1-\alpha)$ percent confidence interval

$$\rho > \frac{N(BMS - F^*\cdot EMS)}{F^*(I\cdot RMS + (IN - I - N)EMS) + N\cdot BMS} = \rho_L\,. \tag{5}$$

An approximate confidence interval bounded above is of the form

$$\rho < \frac{N(F_* \cdot BMS - EMS)}{I\cdot RMS + (IN - I - N)EMS + N\cdot F_*\cdot BMS} = \rho_U\,, \tag{6}$$

where F_* is the upper 100 $(1 - \alpha)$ percentile of the F distribution with ν and $(N - 1)$ degrees of freedom. Approximate two-sided intervals may be derived from (5) and (6) by using the upper 100 $(1 - \alpha/2)$ percentiles.

The accuracy of Satterthwaite's approximation to the distribution of a linear combination of independent mean squares has been studied using computer simulation by, for example, Gaylor and Hopper [1969], and using exact mathematical analysis by, for example, Fleiss [1971]. When, as in the current application, the coefficients of the mean squares are all positive, the approximation has been found to be good.

An important practical use of the lower confidence limit on ρ is in determining the minimum number of raters to employ per subject in a future study in order to assure that their mean ratings have adequate reliability. The Spearman-Brown formula for stepped-up reliability holds in the current model, so that the reliability of the mean of k independent ratings on a subject, say ρ_k, is given by

$$\rho_k = \frac{k\rho}{1 + (k - 1)\rho},$$

where ρ is given by (1). For a given value of k, a 100 $(1 - \alpha)$ percent confidence interval for ρ_k is

$$\rho_k \geq \frac{k\rho_L}{1 + (k - 1)\rho_L},$$

where ρ_L is given in (5). If ρ^* is the minimum acceptable value for the reliability coefficient (e.g., $\rho^* = 0.75$ or 0.80), then the required number of raters per subject should be the smallest integer greater than or equal to

$$k = \frac{\rho^*(1 - \rho_L)}{\rho_L(1 - \rho^*)}. \tag{7}$$

These results are illustrated on data from Winer [1971, p. 288]. The mean squares in Table 2 are for ratings made by $I = 4$ raters on $N = 6$ subjects. Using (2), the reliability of a single rating is estimated to be $\hat{\rho} = 0.74$.

TABLE 2

Mean Squares from an Inter-rater Reliability Study with Four Raters and Six Subjects*

Source	df	Mean Square
Between Subjects	5	24.50
Between Raters	3	5.83
Error	15	1.23

* Data from Winer (1971, p. 288).

The lower 95% confidence bound on ρ may be found as follows. First, the value of F_r, the ratio of the rater to the error mean square, is found (equal to 4.74 in this case). Then, the following process is applied. (i) Calculate the value of ν using (4), with $\hat{\rho}$ replacing ρ; ν is found to equal 11.0 in the current example. (ii) Find the value of F^*, the upper 95th percentile of the F distribution with $N - 1$ and ν degrees of freedom, interpolating if necessary; here, $F^* = 3.20$. (iii) Calculate the value of ρ_L using (5); for these data, finally, $\rho_L = 0.45$.

A similar process is used in finding the upper 95% confidence bound. For $\nu = 11.0$ and $N - 1 = 5$ degrees of freedom, the upper 95th percentile of F is found to be $F^* = 4.71$. Using (6), ρ_U is found to be 0.94.

Suppose that the lower bound of 0.45 is not good enough for the investigator conducting the reliability study, but that the investigator demands, say, 95% assurance that the reliability be at least 0.70. By (7), the minimum value of k is 2.85. Thus, in order to achieve the desired reliability with the desired confidence, the mean of at least three independent ratings should be used.

REFERENCES

Bartko, J. J. The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 1966, *19*, 3–11.

Feldt, L. S. The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 1965, *30*, 357–370.

Fleiss, J. L. On the distribution of a linear combination of independent chi squares. *Journal of the American Statistical Association*, 1971, *66*, 142–144.

Gaylor, D. W. & Hopper, F. N. Estimating the degrees of freedom for linear combinations of mean squares by Satterthwaite's formula. *Technometrics*, 1969, *11*, 691–705.

Satterthwaite, F. E. An approximate distribution of estimates of variance components. *Biometrics*, 1946, *2*, 110–114.

Winer, B. J. *Statistical principles in experimental design*, (2nd ed.). New York: McGraw-Hill, 1971.

Manuscript received 6/13/77
Final version received 10/3/77