Open access • Journal Article • DOI:10.1177/014662169301700302

Estimating Rater Agreement in 2 x 2 Tables: Correction for Chance and Intraclass Correlation. — Source link

Nicole J.-M. Blackman, John J. Koval

Institutions: University of Western Ontario

Published on: 01 Sep 1993 - Applied Psychological Measurement (Sage Publications)

Topics: Cohen's kappa and Intraclass correlation

Estimating Rater Agreement in 2 × 2 Tables: Correction for Chance and Intraclass Correlation

Nicole J-M. Blackman, Glaxo Canada Inc.

John J. Koval, University of Western Ontario

Many estimators of the measure of agreement between two dichotomous ratings of a person have been proposed. The results of Fleiss (1975) are extended, and it is shown that four estimators—Scott's (1955) π coefficient, Cohen's (1960) $\hat{k}$, Maxwell & Pilliner's (1968) r_{11}, and Mak's (1988) $\tilde{\rho}$—are interpretable both as chance-corrected measures of agreement and as intraclass correlation coefficients for different ANOVA models. Relationships among these estimators are established for finite samples. Under Kraemer's (1979) model, it is shown that these estimators are equivalent in large samples, and that the equations for their large sample variances are equivalent. *Index terms: index of agreement, interrater reliability, intraclass correlation, kappa statistic.*

Medicine, epidemiology, psychology, and psychiatry are often interested in classifying people based on a dichotomous outcome. In the absence of a standard against which to assess the quality of their measurements, researchers typically require that a measurement be performed by two raters or by the same rater at two points in time. The degree of agreement between these two ratings is then an indication of the quality of a single measurement.

For the 2 × 2 case—two independent ratings per person based on a dichotomous response—many nonequivalent measures of agreement have been proposed. 2 × 2 agreement indexes have been reviewed in Fleiss (1975), Landis & Koch (1975), and Zwick (1988). Here, four indexes that correct for chance and that are interpretable as intraclass correlation coefficients are investigated.

Notation

Data from a 2 × 2 reliability study can be summarized as in Table 1. Each entry in the table is an observed frequency. Therefore, both raters gave a "+" response n_1 times, Rater 1 gave a "-" response and Rater 2 gave a "+" response n_2 times, and so forth. The marginal totals $n_{.1}$ and $n_{1.}$ indicate that Rater 1 and Rater 2 gave a "+" response with proportions $n_{.1}/n$ and $n_{1.}/n$, respec-

Table 1
Observed Frequencies Resulting From
Classifying n Persons Using
a Dichotomous Outcome

Rater 2	Rater 1 Response		
Response	+	-	Total
+	n_1	n_2	n_1
-	n_3	n_4	n_2
Total	$n_{.1}$	$n_{.2}$	n

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 3, September 1993, pp. 211-223
© Copyright 1993 Applied Psychological Measurement Inc.
0146-6216/93/030211-13$1.90

tively. The marginal totals $n_{.2}$ and $n_{2.}$ indicate that Rater 1 and Rater 2 gave a "-" response with proportions $n_{.2}/n$ and $n_{2.}/n$, respectively.

Agreement Indexes

The simplest agreement index is based on the proportion of persons classified into the same category by the raters. It is given by

$$p_o = \frac{n_1 + n_4}{n} . \tag{1}$$

p_o is known as the "index of crude agreement" (Rogot & Goldberg, 1966) or as the "observed proportion of agreement." However, p_o does not account for the level of agreement expected by chance alone when the two ratings are independent. As discussed in Fleiss (1975), p_o can be suitably corrected for chance in the following manner. Let p_e denote the expected value of p_o when there is no agreement other than by chance. Then $(p_o - p_e)$ represents agreement beyond chance, and $(1 - p_e)$ is the maximum attainable amount of agreement beyond chance. The ratio of these differences, denoted A is given by

$$A = \frac{p_o - p_e}{1 - p_e} . \tag{2}$$

A is a standardized, chance-corrected measure of agreement with the following properties. If there is perfect agreement, then $A = 1$. If observed agreement is equal to expected agreement, then $A = 0$. The minimum value of A is equal to $-p_e/(1 - p_e)$. If the marginal probabilities are such that $p_e = .5$, then the minimum is equal to -1; otherwise, it is between -1 and 0.

Scott's (1955) π coefficient, Cohen's (1960) $\hat{k}$, and Mak's (1988) $\tilde{\rho}$ are indexes similar in form to A. They differ only in their definition of proportion of agreement by chance, p_e. Maxwell & Pilliner's (1968) r_{11} is not in the form of A but does possess the same properties as estimators similar to A.

In proposing the π coefficient, Scott (1955) assumed marginal homogeneity as well as independence; that is, both raters have the same probability of giving a "+" response. Thus, Scott's definition of expected proportion of agreement by chance, $p_e(\pi)$, is equal to $\bar{p}^2 + \bar{q}^2$, where

$$\bar{p} = \frac{2n_1 + n_2 + n_3}{2n} \tag{3}$$

and

$$\bar{q} = \frac{2n_4 + n_2 + n_3}{2n} . \tag{4}$$

Scott's π is therefore given by

$$\pi = \frac{4(n_1 n_4 - n_2 n_3) - (n_2 - n_3)^2}{(2n_1 + n_2 + n_3)(2n_4 + n_2 + n_3)} . \tag{5}$$

In proposing the kappa statistic $\hat{k}$, Cohen assumed only independence. Cohen's definition of the expected proportion of agreement by chance, $p_e(\hat{k})$, is equal to $(n_{1.}/n)(n_{.1}/n) + (n_{2.}/n)(n_{.2}/n)$. Cohen's $\hat{k}$ is given by

$$\hat{k} = \frac{2(n_1n_4 - n_2n_3)}{(n_1 + n_2)(n_2 + n_4) + (n_1 + n_3)(n_3 + n_4)} . \tag{6}$$

Mak (1988) proposed an agreement measure applicable to the case of two or more raters with a dichotomous outcome. For the 2 × 2 case, his value for chance agreement is obtained as follows. First, select any two individuals and for each individual select a rater. Then ask, "What is the probability that the responses of these two raters will be the same?" If all possible pairs of individuals are used, Mak's expected proportion by chance, $p_e(\tilde{\rho})$, is given by

$$p_e(\tilde{p}) = 1 - \frac{1}{2n(n-1)} [(2n_1 + n_2 + n_3)(2n_4 + n_2 + n_3) - (n_2 + n_3)] , \tag{7}$$

which is simply the probability that the raters' responses will differ for all persons, less the probability that the responses will be different for the same person, subtracted from 1. Mak's estimator $\tilde{\rho}$ is thus given by

$$\tilde{\rho} = \frac{4n_1n_4 - (n_2 + n_3)^2 + (n_2 + n_3)}{(2n_1 + n_2 + n_3)(2n_4 + n_2 + n_3) - (n_2 + n_3)} . \tag{8}$$

The agreement measure proposed by Maxwell & Pilliner (1968), denoted r_{11}, is given by

$$r_{11} = \frac{2(n_1n_4 - n_2n_3)}{(n_1 + n_2)(n_3 + n_4) + (n_1 + n_3)(n_2 + n_4)} . \tag{9}$$

If p_e denotes the expected value of p_o assuming only independence (and not marginal homogeneity) and M denotes the arithmetic mean, then

$$r_{11} = \frac{p_o - p_e}{2M} . \tag{10}$$

Hence r_{11} is a measure of agreement standardized not by the maximum possible amount of beyond chance agreement but by the mean of the raters' variances. These results extend those given in Fleiss (1975) by including Mak's $\tilde{\rho}$ as a chance-corrected estimator.

Intraclass Correlation

As discussed in Fleiss (1975) and Landis & Koch (1975), ANOVA procedures can be applied to dichotomous data to obtain estimates of various intraclass correlation coefficients (ICRs) according to a specified model. One of several ANOVA models (e.g., one-way random, two-way random, two-way mixed) may be selected depending on how the data were collected and what inferences are to be made. Table 2 gives a schematic representation of two independent measurements taken on a random sample of n persons.

Let $X_{ij} = 0$ for a "-" response and 1 for a "+" response. Then, let

$$S_1 = X_{..}^2/2n = (2n_1 + n_2 + n_3)^2/2n , \tag{11}$$

$$S_2 = \sum_i \sum_j X_{ij}^2 = (2n_1 + n_2 + n_3) , \tag{12}$$

$$S_3 = \sum_j X_{.j}^2/n = [(n_1 + n_3)^2 + (n_1 + n_2)^2]/n , \tag{13}$$

Table 2
Responses (X_{ij}) for Two Measurements on Each of n Persons

Examinee	Response Measurement 1	 Measurement 2	Total
1	X_{11}	X_{12}	$X_{1.}$
2	X_{21}	X_{22}	X_{2}
3	X_{31}	X_{32}	X_{3}
.	.	.	.
.	.	.	.
.	.	.	.
n	X_{n1}	X_{n2}	X_{n}
Total	X_{1}	X_{2}	$X_{.}$

and

$$S_4 = \sum_i X_{i.}^2/2 = (4n_1 + n_2 + n_3)/2 \; . \tag{14}$$

Let $\mathrm{SS_b}$, $\mathrm{SS_w}$, $\mathrm{SS_j}$, and $\mathrm{SS_r}$, denote the between persons, within persons, between raters, and residual sum of squares, respectively. Thus

$$\mathrm{SS_b} = \mathrm{S_4} - \mathrm{S_1} = 4n_1 + n_2 + n_3/2 - (2n_1 + n_2 + n_3)^2/2n = [4n_1n_4 + (n_1 + n_4)(n_2 + n_3)]/2n \; , \tag{15}$$

$$\mathrm{SS_w} = \mathrm{S_2} - \mathrm{S_4} = (2n_1 + n_2 + n_3) - (4n_1 + n_2 + n_3)/2 = (n_2 + n_3)/2 \; , \tag{16}$$

$$\mathrm{SS_j} = \mathrm{S_3} - \mathrm{S_1} = \frac{(n_1 + n_2)^2 + (n_1 + n_3)^2}{n} - \frac{(2n_1 + n_2 + n_3)^2}{2n} = \frac{(n_2 - n_3)^2}{2n} \; , \tag{17}$$

and

$$\mathrm{SS_r} = \mathrm{S_1} + \mathrm{S_2} - \mathrm{S_3} - \mathrm{S_4} = \frac{4n_2n_3 + (n_1 + n_4)(n_2 + n_3)}{2n} \; , \tag{18}$$

with $n - 1$, n, 1, and $n - 1$ degrees of freedom, respectively. The total sum of squares is

$$\mathrm{SS_T} = [4n_1n_4 + (n_1 + n_4)(n_2 + n_3)]/2n(n - 1) + (n_2 + n_3)/2n \tag{19}$$

with $2n - 1$ degrees of freedom.

If potential differences in raters' means are ignored, then a simple one-way random effects model is used—variation between persons and variation within persons. The ICR is the amount of the total variability that is explained by the within person variation. The appropriate estimate of the ICR is given by (Bartko, 1966)

$$R_1 = \frac{\mathrm{MS_b} - \mathrm{MS_w}}{\mathrm{MS_b} + \mathrm{MS_w}} = \frac{[4n_1n_4 + (n_1 + n_4)(n_2 + n_3)]/2n(n - 1) - (n_2 + n_3)/2n}{[4n_1n_4 + (n_1 + n_4)(n_2 + n_3)]/2n(n - 1) + (n_2 + n_3)/2n}$$

$$= \frac{4(n_1n_4 - n_2n_3) - (n_2 + n_3)^2 + (n_2 + n_3)}{(2n_1 + n_2 + n_3)(2n_4 + n_2 + n_3) - (n_2 + n_3)} \; , \tag{20}$$

which is Mak's $\tilde{\rho}$. If n is sufficiently large so that the difference between n and $(n - 1)$ is negligible, then R_1 can be approximated by

$$R_2 = \frac{[4n_1n_4 + (n_1 + n_4)(n_2 + n_3)]/2n^2 - (n_2 + n_3)/2n}{[4n_1n_4 + (n_1 + n_4)(n_2 + n_3)]/2n^2 + (n_2 + n_3)/2n} = \frac{4(n_1n_4 - n_2n_3) - (n_2 - n_3)^2}{(2n_1 + n_2 + n_3)(2n_4 + n_2 + n_3)} , \tag{21}$$

which is Scott's π coefficient.

Suppose the raters are considered to be a random sample from a larger population of potential raters. This is a two-way random effects model. Then, the appropriate estimate of the ICR is given by (Bartko, 1966)

$$R_3 = \frac{(\mathrm{MS_b} - \mathrm{MS_r}/2)}{(\mathrm{MS_b} + \mathrm{MS_r})/2 + (\mathrm{MS_j} - \mathrm{MS_r})/n + \mathrm{MS_r}}$$

$$= \frac{\dfrac{4n_1n_4 + (n_1 + n_4)(n_2 + n_3)}{4n(n-1)} - \dfrac{4n_2n_3 + (n_1 + n_4)(n_2 + n_3)}{4n(n-1)}}{\dfrac{4n_2n_3 + (n_1 + n_4)(n_2 + n_3)}{2n(n-1)} + \dfrac{(n_2 - n_3)^2}{2n^2} - \dfrac{4n_2n_3 + (n_1 + n_4)(n_2 + n_3)}{2n^2(n-1)}} , \tag{22}$$

where $\mathrm{MS_b}$ = mean square between persons, $\mathrm{MS_r}$ = mean square residual, and $\mathrm{MS_j}$ = mean square between raters.

Again, suppose that n is sufficiently large so that n is effectively equal to $(n - 1)$. Then

$$R_3 = \frac{\dfrac{1}{n^2}(n_1n_4 - n_2n_3)}{\dfrac{1}{2n^2}[2n_1n_4 + (n_1 + n_4)(n_2 + n_3) + n_2^2 + n_3^2] - \dfrac{1}{2n^3}[4n_2n_3 + (n_1 + n_4)(n_2 + n_3)]} . \tag{23}$$

If terms of order $1/n$ are ignored,

$$R_3 = \frac{2(n_1n_4 - n_2n_3)}{(n_1 + n_2)(n_2 + n_4) + (n_1 + n_3)(n_2 + n_3)} , \tag{24}$$

which is Cohen's $\hat{k}$.

If the raters are considered to be a fixed set, then a two-way mixed effects model would be used. The appropriate ICR then is estimated by (Bartko, 1966)

$$R_4 = \frac{\mathrm{MS_b} - \mathrm{MS_r}}{\mathrm{MS_b} + \mathrm{MS_r}} = \frac{4n_1n_4 + (n_1 + n_4)(n_2 + n_3) - [4n_2n_3 + (n_1 + n_4)(n_2 + n_3)]}{4n_1n_4 + (n_1 + n_4)(n_2 + n_3) + [4n_2n_3 + (n_1 + n_4)(n_2 + n_3)]}$$

$$= \frac{2(n_1n_4 - n_2n_3)}{(n_1 + n_2)(n_3 + n_4) + (n_1 + n_3)(n_2 + n_4)} , \tag{25}$$

which is Maxwell and Pilliner's r_{11}.

Thus, Scott's π, Cohen's $\hat{k}$, Mak's $\tilde{\rho}$, and Maxwell and Pilliner's r_{11} are interpretable both as chance-corrected measures and as ICRs. Correspondence between definitions of expected proportion due to chance and assumptions on rater effects are presented in Table 3. These results extend Fleiss (1975) by including Mak's $\tilde{\rho}$ and describing the estimators in terms of traditional ANOVA models.

Finite Sample Relationships

In finite samples, the following relationships hold:

$$|r_{11}| \geq |\hat{k}| , \tag{26}$$

Table 3
Expected Proportion Due to Chance (p_e) for ICR Models

Index	Definition of p_e	ANOVA Model
π	$\left(\frac{2n_1 + n_2 + n_3}{2n}\right)^2 + \left(\frac{2n_4 + n_2 + n_3}{2n}\right)^2$	Asymptotically one-way random effects
$\tilde{\rho}$	$1 - \frac{[(2n_1 + n_2 + n_3)(2n_4 + n_2 + n_3) - (n_2 + n_3)]}{2n(n - 1)}$	One-way random effects
$\hat{k}$	$\frac{1}{n^2}[(n_1 + n_2)(n_1 + n_3) + (n_2 + n_4)(n_3 + n_4)]$	Asymptotically two-way random effects
r_{11}	$\frac{1}{n^2}[(n_1 + n_2)(n_1 + n_3) + (n_2 + n_4)(n_3 + n_4)]$	Two-way mixed effects

$$\hat{k} \geq \pi \ , \tag{27}$$

$$\tilde{\rho} \geq \pi \ , \tag{28}$$

and

$$r_{11} \geq \pi \ . \tag{29}$$

Proof:

From Equation 26,

$$\begin{aligned} |r_{11}| &= \left| \frac{2(n_1 n_4 - n_2 n_3)}{(n_1 + n_2)(n_3 + n_4) + (n_1 + n_3)(n_2 + n_4)} \right| \\ &= \left| \frac{2(n_1 n_4 - n_2 n_3)}{(n_1 + n_2)(n_2 + n_4) + (n_1 + n_3)(n_3 + n_4) - (n_2 - n_3)^2} \right| \\ &\geq \left| \frac{2(n_1 n_4 - n_2 n_3)}{(n_1 + n_2)(n_2 + n_4) + (n_1 + n_3)(n_3 + n_4)} \right| = |\hat{k}| \ . \end{aligned} \tag{30}$$

Equality holds when $n_2 = n_3$ (Fleiss, 1975, p. 658).

Table 3 shows that

$$p_e(\hat{k}) = p_e(\pi) - \frac{1}{2n^2}(n_2 - n_3)^2 \ . \tag{31}$$

Hence,

$$\hat{k} = \frac{p_o - p_e(\pi) + \frac{1}{2n^2}(n_2 - n_3)^2}{1 - p_e(\pi) + \frac{1}{2n^2}(n_2 - n_3)^2} \geq \frac{p_o - p_e(\pi)}{1 - p_e(\pi)} = \pi \ . \tag{32}$$

Equality holds when $n_2 = n_3$.

It follows from Table 3 that

$$p_e(\tilde{\rho}) = p_e(\pi) - \frac{1}{2n^2(n-1)}\,[4n_1n_4 + (n_1 + n_4)(n_2 + n_3)]\ . \tag{33}$$

Hence, $\tilde{\rho}$ can be expressed as

$$\tilde{\rho} = \frac{p_o - p_e(\pi) + \dfrac{1}{2n^2(n-1)}\,[4n_1n_4 + (n_1 + n_4)(n_2 + n_3)]}{1 - p_e(\pi) + \dfrac{1}{2n^2(n-1)}\,[4n_1n_4 + (n_1 + n_4)(n_2 + n_3)]} \geq \frac{p_o - p_e(\pi)}{1 - p_e(\pi)} = \pi\ . \tag{34}$$

From Equation 26,

$$r_{\text{II}} = \frac{2(n_1n_4 - n_2n_3)}{(n_1 + n_2)(n_3 + n_4) + (n_1 + n_3)(n_2 + n_4)} = \frac{4(n_1n_4 - n_2n_3)}{2(n_1 + n_2)(n_3 + n_4) + 2(n_1 + n_3)(n_2 + n_4)}$$

$$= \frac{4(n_1n_4 - n_2n_3)}{(2n_1 + n_2 + n_3)(2n_4 + n_2 + n_3) - (n_2 - n_3)^2} \geq \frac{4(n_1n_4 - n_2n_3)}{(2n_1 + n_2 + n_3)(2n_4 + n_2 + n_3)} = \pi\ . \tag{35}$$

Equality holds when $n_2 = n_3$.

Asymptotic Relationships

Bloch & Kraemer (1989) proposed a population model for 2 × 2 tables. It is a simplification of Kraemer's (1979) model and Mak's (1988) model. Mak proposed the more general model to deal with the analysis of measurements on animals in a litter with a dichotomous response variable (hence, variable number of responses per group. This would be equivalent to having a variable number of raters per person. The "simple" case is two raters per person and two animals per litter). The derivation of the model for 2 × 2 tables is as follows. Let X_1 and X_2 be dichotomous response variables representing the scores of two raters on one person. Let 0 represent a "–" response and 1 a "+" response. Hence,

$$\Pr(X_i = 1) = P,\ i = 1,2\ , \tag{36}$$

$$\Pr(X_i = 0) = Q = 1 - P,\ i = 1,2\ , \tag{37}$$

$$\mathrm{E}(X_i) = P,\ i = 1,2\ , \tag{38}$$

and

$$\mathrm{Var}(X_i) = PQ,\ i = 1,2\ . \tag{39}$$

Note that this model assumes marginal homogeneity. Define the intraclass κ as

$$\kappa = \frac{\mathrm{cov}(X_1,X_2)}{[\mathrm{Var}(X_1)\mathrm{Var}(X_2)]^{1/2}} = \frac{\mathrm{cov}(X_1,X_2)}{\mathrm{Var}(X_1)} = \frac{\mathrm{cov}(X_1,X_2)}{PQ}\ , \tag{40}$$

where $\mathrm{cov}(X_1,X_2)$ is the covariance between X_1 and X_2. As shown in Appendix A, this yields the common correlation model given in Table 4.

Table 4
The Common Correlation Model for Two Correlated Dichotomous Outcomes

Rater 2	Rater 1 Response		
Response	+	-	Total
+	$P^2 + \kappa PQ$	$(1 - \kappa)PQ$	P
-	$(1 - \kappa)PQ$	$Q^2 + \kappa PQ$	Q
Total	P	Q	1

Under this common correlation model, as $n \rightarrow \infty$,

$$n_1 \rightarrow n(P^2 + \kappa PQ) \ , \tag{41}$$

$$n_2, n_3, \rightarrow n[(1 - \kappa)PQ] \ , \tag{42}$$

and

$$n_4 \rightarrow n(Q^2 + \kappa PQ) \ . \tag{43}$$

Bloch & Kraemer (1989) used the common correlation model to compute the maximum likelihood estimates $\hat{P}$ and $\hat{\kappa}_I$ of P and κ as

$$\hat{P} = \frac{2n_1 + n_2 + n_3}{2n} \tag{44}$$

and

$$\hat{\kappa}_I = \frac{4(n_1n_4 - n_2n_3) - (n_2 - n_3)^2}{(2n_1 + n_2 + n_3)(2n_4 + n_2 + n_3)} \ . \tag{45}$$

They noted the equivalence of $\hat{\kappa}_I$ to Scott's (1955) π. It will now be shown that Scott's π, Cohen's $\hat{k}$, Mak's $\tilde{\rho}$, and Maxwell & Pilliner's r_{11} are equivalent in large samples when outcomes are generated by Bloch & Kraemer's common correlation model.

1. Claim that

$$r_{11} \rightarrow \hat{k} \text{ as } n \rightarrow \infty \ . \tag{46}$$

Consider the following proof:

$$\left| \frac{r_{11}}{\hat{k}} \right| = \left| \frac{(n_1 + n_2)(n_2 + n_4) + (n_1 + n_3)(n_3 + n_4)}{(n_1 + n_2)(n_2 + n_4) + (n_1 + n_3)(n_3 + n_4) - (n_2 - n_3)^2} \right|$$

$$\rightarrow \frac{(n_1 + n_2)(n_2 + n_4) + (n_1 + n_3)(n_3 + n_4)}{(n_1 + n_2)(n_2 + n_4) + (n_1 + n_3)(n_3 + n_4)} = 1 \quad \text{as } n \rightarrow \infty \ . \tag{47}$$

2. Claim that

$$\hat{k} \rightarrow \pi \text{ as } n \rightarrow \infty \ . \tag{48}$$

$$p_e(\pi) - p_e(\hat{k}) = \frac{1}{2n^2} (n_2 - n_3)^2 \rightarrow 0 \text{ as } n \rightarrow \infty \tag{49}$$

which gives

$$|\hat{\kappa} - \pi| = \left| \frac{p_o - p_e(\hat{\kappa})}{1 - p_e(\hat{\kappa})} - \frac{p_o - p_e(\pi)}{1 - p_e(\pi)} \right| = \left| \frac{[p_e(\hat{\kappa}) - p_e(\pi)](p_o - 1)}{[1 - p_e(\pi)][1 - p_e(\pi)]} \right| \rightarrow 0 \text{ as } n \rightarrow \infty \ . \tag{50}$$

Equations 46 and 48 together imply

$$r_{11} \rightarrow \pi \ \textit{as} \ n \rightarrow \infty \ . \tag{51}$$

3. Claim that

$$\tilde{p} \rightarrow \pi \ \textit{as} \ n \rightarrow \infty \ . \tag{52}$$

$$p_e(\pi) - p_e(\tilde{\rho}) = \frac{1}{2n^2(n - 1)} [4n_1n_4 + (n_1 + n_4)(n_2 + n_3)] \rightarrow 0 \text{ as } n \rightarrow \infty \ . \tag{53}$$

It follows then that

$$|\tilde{\rho} - \pi| = \left| \frac{p_o - p_e(\tilde{\rho})}{1 - p_e(\tilde{\rho})} - \frac{p_o - p_e(\pi)}{1 - p_e(\pi)} \right| = \left| \frac{[p_e(\tilde{\rho}) - p_e(\pi)](p_o - 1)}{[1 - p_e(\pi)][1 - p_e(\tilde{\rho})]} \right| \rightarrow 0 \text{ as } n \rightarrow \infty \ . \tag{54}$$

Thus, $\hat{k}$, r_{11}, and $\tilde{\rho}$ are asymptotically equivalent to Scott's π, which is also the maximum likelihood estimator of κ under the conditions of the common correlation model. Because the probability distribution of the common correlation model satisfies the usual regularity conditions, the maximum likelihood estimator is consistent (Cox & Hinkley, 1974, p. 281). Therefore $\hat{k}$, r_{11}, and $\tilde{\rho}$ are consistent.

Asymptotic Variance

Using the following result due to Fisher (1970, p. 311) that is based on a first-order Taylor series expansion,

$$\frac{1}{n} \operatorname{Var}(\hat{\kappa}) = \left[\sum_{i=1}^{k} w_i \left(\frac{d\hat{\kappa}}{dn_i} \right)^2 \right] , \tag{55}$$

where

$$w_1 = P^2 + \kappa PQ , \tag{56}$$

$$w_2, w_3 = PQ(1 - \kappa) , \tag{57}$$

and

$$w_4 = Q^2 + \kappa PQ , \tag{58}$$

and $d\hat{\kappa}/dn_i$ is the first partial derivative of the estimator of $\hat{\kappa}$ with respect to n_i. Bloch & Kraemer (1989) derived the variance of the maximum likelihood estimator $\hat{\kappa}$ as

$$\mathrm{Var}(\hat{\kappa}) = \frac{(1-\kappa)}{n}\left[(1-\kappa)(1-2\kappa) + \frac{\kappa(2-\kappa)}{2PQ}\right] . \tag{59}$$

When $\hat{\kappa}$ is set equal to Cohen's $\hat{k}$, Mak's $\tilde{\rho}$, or Maxwell-Pilliner's r_{11}, then Equations 55–58 yield identical asymptotic variances (Equation 59) for $\hat{k}$, $\tilde{\rho}$, and r_{11}. This is expected because the estimators are asymptotically equivalent.

Mak (1988) used a different approach, based on first-order Taylor series expansions, to obtain the asymptotic variance of $\tilde{\rho}$. For the 2 × 2 case, Mak's formula reduces to Equation 59 (see Appendix B).

The Estimated Variance of Cohen's $\hat{k}$

Using the results of Rao (1965, p. 321), Fleiss, Cohen, & Everitt (1969) derived an estimate for the large sample variance of Cohen's kappa. Let

$$p_{1.} = n_{1.}/n \quad p_{.1} = n_{.1}/n \quad p_{2.} = n_{2.}/n \quad p_{.2} = n_{.2}/n$$

$$p_{11} = n_1/n \quad p_{12} = n_2/n \quad p_{21} = n_3/n \quad p_{22} = n_4/n \ , \tag{60}$$

where $n_{.1}$, $n_{.2}$, $n_{1.}$, and $n_{2.}$ are defined as in Table 1. Let p_e denote Cohen's definition of chance agreement as shown in Table 3.

Then

$$\mathrm{Var}(\hat{k}) = \frac{A + B - C}{(1-p_e)^2 n} \ , \tag{61}$$

where

$$A = \sum_{i=1}^{2} p_{ii}[1 - (p_{i.} + p_{.i})(1-\hat{k})]^2 \ , \tag{62}$$

$$B = (1-\hat{k})^2 \sum \sum_{i \neq j} p_{ij}(p_{i.} + p_{.j})^2 \ , \tag{63}$$

and

$$C = [\hat{k} - p_e(1-\hat{k})^2] \ . \tag{64}$$

Asymptotically (at $n_i = nw_i$), Equation 61 also reduces to Equation 59. Thus, all three approaches lead to the same asymptotic variance formula.

Discussion

When a sample of persons is rated on a quantitative scale, reliability is traditionally measured by the ICR. For ratings on categorical—specifically, dichotomous—scales, reliability has been measured in terms of beyond chance agreement. Fleiss (1975) showed that π, $\hat{k}$, and r_{11} are chance-corrected measures and intraclass coefficients, but advocated the use of only $\hat{k}$ and r_{11}. The formulation of π assumes homogeneous rater marginals—an assumption that Fleiss (1975) felt to be unreasonable. The results of Fleiss (1975) have been extended here to include Mak's $\tilde{\rho}$; in terms of an ICR, $\tilde{\rho}$ is the exact version of π.

In finite samples, π, $\hat{k}$, r_{11}, and $\tilde{\rho}$ differ in well-defined ways. Blackman (1991) compared the moments of the distributions of these estimators in small samples. Which index to use depends on the definition of chance that is considered appropriate, or on the assumptions made about rater effects.

Kraemer (1979) and Bloch & Kraemer (1989) described the difference between indexes of agreement and indexes of association and formulated a well-defined population model for agreement. π, $\hat{k}$, r_{11}, and $\tilde{\rho}$—indexes of agreement in 2 × 2 tables under this model—have been shown to be asymptotically equivalent to each other. All are consistent estimators of the true index of rater agreement.

The asymptotic variance formula has been derived in three different ways. The accuracy of this formula in small samples is described in Blackman (1991).

Appendix A

Assume that the marginal distributions of X_1 and X_2 are identical, but that X_1 and X_2 are correlated. Hence

$$P = \Pr(X_i = 1),\ i = 1,2\ , \tag{65}$$

$$Q = \Pr(X_i = 0) = 1 - P,\ i = 1,2\ , \tag{66}$$

$$p_{jk} = \Pr(X_1 = j,\ X_2 = k),\ j,k = 0,1\ , \tag{67}$$

and

$$\kappa = \mathrm{corr}(X_1,X_2)\ . \tag{68}$$

But κ also may be written as

$$\frac{\mathrm{cov}(X_1,X_2)}{\mathrm{Var}(X_1)} = \frac{\mathrm{E}(X_1X_2) - [\mathrm{E}(X_1)\mathrm{E}(X_2)]}{\mathrm{E}(X_1^2) - \mathrm{E}(X_1)^2} = \frac{p_{11} - P^2}{PQ}\ . \tag{69}$$

Solving for p_{11},

$$p_{11} = P^2 + \kappa PQ\ . \tag{70}$$

Moreover, because

$$p_{11} + p_{01} = P\ , \tag{71}$$

then

$$p_{01} = (1 - \kappa)PQ\ . \tag{72}$$

Similarly for p_{10}. Finally, because

$$p_{00} + p_{01} + p_{10} + p_{11} = 1\ , \tag{73}$$

then

$$p_{00} = Q^2 + \kappa PQ\ . \tag{74}$$

Appendix B

According to Mak (1988, p. 348), the variance of $\tilde{\rho}$ is given by

$$\mathrm{Var}(\tilde{\rho}) = \frac{1}{n}(d_1,d_2)\mathbf{W}(d_1,d_2)'\ , \tag{75}$$

where

$$d_1 = \frac{1}{-[P(1 - P)]} , \tag{76}$$

$$d_2 = \frac{(1 - 2P)(1 - \kappa)}{[P(1 - P)]} , \tag{77}$$

$$\mathbf{W} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(\mathrm{Z}_i\mathrm{Z}_i') - \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(\mathrm{Z}_i)\mathrm{E}(\mathrm{Z}_i)' , \tag{78}$$

and

$$Z_i = \frac{1}{2} \begin{pmatrix} R_i(2 - R_i) \\ R_i \end{pmatrix} , \tag{79}$$

and R_i is the sum of the two ratings for person i. Then

$$\mathrm{E}(\mathrm{Z}_i) = \begin{pmatrix} P(1 - P)(1 - \kappa) \\ P \end{pmatrix} \tag{80}$$

and

$$\mathrm{E}(\mathrm{Z}_i\mathrm{Z}_i') = \frac{1}{4} \begin{pmatrix} n_2 + n_3 & n_2 + n_3 \\ n_2 + n_3 & 4n_1 + n_2 + n_3 \end{pmatrix} . \tag{81}$$

Hence

$$\begin{aligned} \mathrm{Var}(\tilde{\rho}) &= \frac{(1 - \kappa)}{2nP(1 - P)} \{1 - (1 - \kappa)[(1 - 2P)^2(1 - \kappa) + 2P(1 - P)]\} \\ &= \frac{(1 - \mathrm{k})}{n} \left\{ \frac{2P(1 - P)[2(1 - \kappa)^2 - (1 - \kappa)]}{2P(1 - P)} + \frac{1 - 1 + 2\kappa - \kappa^2}{2P(1 - P)} \right\} \\ &= \frac{(1 - \kappa)}{n} \left\{ (1 - \kappa)(1 - 2\kappa) + \frac{\kappa(2 - \kappa)}{2P(1 - P)} \right\} . \end{aligned} \tag{82}$$

References

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

Blackman, N. J-M. (1991). *A comparison of measures of agreement between two dichotomous ratings.* Unpublished M.Sc. thesis, University of Western Ontario, London, Canada.

Bloch, D. A., & Kraemer, H. C. (1989). 2 × 2 Kappa coefficients: Measures of agreement or association. *Biometrics, 45,* 269-287.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics.* London: Wiley.

Fisher, R. A. (1970). *Statistical methods for research workers* (14th ed.). New York: Hafner.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31,* 651-659.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 5,* 323–327.

Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika, 44,* 461–472.

Landis, R. J., & Koch, G. G. (1975). A review of statistical methods in the analysis of data arising from observer reliability studies (Parts I and II). *Statistica Neerlandica, 3,* 101–161.

Mak, T. K. (1988). Analyzing intraclass correlation for dichotomous variables. *Applied Statistics, 37,* 344–352.

Maxwell, A. E., & Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. *The British Journal of Mathematical and Statistical Psychology, 21,* 105–116.

Rao, C. R. (1965). *Linear statistical inference and its applications* (1st ed.). New York: Wiley.

Rogot, E., & Goldberg, I. D. (1966). A proposed index for measuring agreement in test-retest studies. *Journal of Chronic Diseases, 19,* 991–1006.

Scott, W. A. (1955). Reliability or content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19,* 321–325.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103,* 374–378.

Acknowledgments

This work was partially supported by an operating grant from the National Sciences and Engineering Research Council (Canada).

Author's Address

Send requests for reprints or further information to John J. Koval, Department of Epidemiology and Biostatistics, Kresge Building, The University of Western Ontario, London, Ontario, Canada, N6A 5C1. Internet: jkoval@biostats.uwo.ca.