The Reliability of Dichotomous Judgments: Unequal Numbers of Judges per Subject

Joseph L. Fleiss
Columbia University and New York State Psychiatric Institute

Jack Cuzick
Columbia University

Consider a reliability study in which different subjects are judged on a dichotomous trait by different sets of judges, possibly unequal in number. A kappa-like measure of reliability is proposed, its correspondence to an intraclass correlation coefficient is pointed out, and a test for its statistical significance is presented. A numerical example is given.

Following Cohen's (1960) development of kappa as a chance-corrected measure of agreement between a fixed pair of judges on a nominal scale, several authors have generalized kappa to the case of more than two judges. Landis and Koch (1977a) and Light (1971) considered the case of a fixed set of judges, whereas Fleiss (1971) and Landis and Koch (1977b) considered the case of varying sets of judges. In this paper a reliability study is considered in which different subjects are judged by different judges and the number of judges per subject varies. Attention is restricted to the case of dichotomous judgments.

For example, the subjects may be hospitalized mental patients, the studied characteristic may be the presence or absence of schizophrenia, and the judges may be those psychiatry residents, out of a much larger pool, who happen to be on call when a patient is newly admitted. Not only may the particular residents responsible for one patient be different from those responsible for another, but different numbers of residents may provide diagnoses on different patients.

In this paper, a kappa-like measure of reliability is proposed as appropriate to this kind of study. The measure varies from negative values for the case of less than chance agreement, through zero for a degree of agreement exactly in accordance with chance, to unity for perfect agreement. Its correspondence to an intraclass correlation coefficient is pointed out, and a test for its significance is presented.

The Measure of Similarity

Let N denote the number of subjects under study, n_i the number of judges judging the i^{th} subject, and x_i the number of positive judgments on the i^{th} subject. Define $p_i = x_i/n_i$ and $q_i = 1-p_i$. Further, define

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 4 Fall 1979 pp. 537–542

$$\bar{n} = \frac{1}{N} \Sigma n_i , \qquad [1]$$

as the mean number of judges per subject,

$$\bar{p} = \frac{1}{N\bar{n}} \Sigma x_i , \qquad [2]$$

as the overall proportion of positive judgments, and $\bar{q} = 1-\bar{p}$. In these and all subsequent equations, summation is over the N subjects. The proposed measure is

$$\kappa = 1 - \frac{\Sigma n_i p_i q_i}{N(\bar{n}-1)\bar{p}\,\bar{q}} . \qquad [3]$$

When the number of judges per subject is constant, this statistic is identical to the version of kappa proposed earlier by Fleiss (1971).

The statistic given in Equation 3 has the following properties:

1. If there is no intersubject variation in the proportion of positive judgments (i.e., if $p_i = \bar{p}$ for all i), then there is less agreement (or more disagreement) among the judgments within than between the N subjects. In this case κ may be seen to assume its minimum value of $-1/(\bar{n}-1)$.
2. If the several proportions p_i vary exactly as binomial proportions with parameters n_i and a common probability $\bar{p}$, then there is as much agreement among the judgments within the subjects as there is between the subjects. In this case, the value of κ is equal to 0.
3. If each proportion p_i assumes either the values 0 or 1, then there is perfect agreement among the judgments on each subject. In this case, κ may be seen to assume the value 1.

Kappa as an Approximate Intraclass Correlation Coefficient

Landis and Koch (1977b) approached the measurement of reliability for varying sets of judges by applying the algebra of a one-way random effects analysis of variance (Snedecor & Cochran, 1967) to the data obtained by coding positive judgments "1" and negative judgments "0". In the present notation, the mean square between subjects (BMS) is

$$\mathrm{BMS} = \frac{1}{N-1} \Sigma n_i (p_i - \bar{p})^2 \qquad [4]$$

and the mean square within subjects (WMS) is

$$\mathrm{WMS} = \frac{1}{N(\bar{n}-1)} \Sigma n_i p_i q_i . \qquad [5]$$

They proposed as a measure of reliability the sample intraclass correlation coefficient,

$$r = \frac{\mathrm{BMS} - \mathrm{WMS}}{\mathrm{BMS} + (n_o - 1)\mathrm{WMS}} , \qquad [6]$$

where

$$n_o = \bar{n} - \frac{s_n^2}{N\bar{n}} , \qquad [7]$$

is the constant used to estimate "true score" variance in a one-way layout, and

$$S_n^2 = \frac{1}{N-1} \Sigma (n_i - \bar{n})^2 \quad , \qquad [8]$$

is the variance of the sample sizes.

It may be easily checked that if BMS is redefined to have a divisor N instead of $N-1$, then

$$r = \frac{\kappa}{1-f} \quad , \qquad [9]$$

where

$$f = \frac{S_n^2}{N\bar{n}^2} (1-\kappa) . \qquad [10]$$

Thus, if N is even moderately large, Landis and Koch's measure is virtually identical to κ. This represents yet another correspondence between kappa statistics and intraclass correlation coefficients (Fleiss, 1975; Fleiss & Cohen, 1973). Given this correspondence, the degree of reliability may be described as good if κ exceeds approximately .60 (Landis & Koch, 1977a).

The Distribution of Kappa

Consider testing the hypothesis that the underlying probability of a positive judgment for each subject is constant. A test statistic for this hypothesis is

$$\chi^2 = \frac{\Sigma n_i (p_i - \bar{p})^2}{\bar{p}\,\bar{q}} \quad . \qquad [11]$$

When the hypothesis is true and N is large, χ^2 is approximately normally distributed with mean $N-1$ and variance

$$\mathrm{Var}(\chi^2) \sim \frac{2N(\bar{n}_H - 1)}{\bar{n}_H} + \frac{N(\bar{n} - \bar{n}_H)(1 - 4\bar{p}\,\bar{q})}{\bar{n}\,\bar{n}_H\,\bar{p}\,\bar{q}} \qquad [12]$$

(Cuzick, 1977; Haldane, 1939), where $\bar{n}_H$ is the harmonic mean of the number of judges,

$$\bar{n}_H = \frac{N}{\Sigma \frac{1}{n_i}} \quad . \qquad [13]$$

Given the easily verified identity

$$\kappa = \frac{\chi^2 - N}{N(\bar{n}-1)} \quad , \qquad [14]$$

it follows that, when N is large and the N subjects have the same underlying probability of a positive judgment, κ is approximately normally distributed with mean

$$E(\kappa) \sim \frac{-1}{N(\bar{n}-1)} \sim 0 \qquad [15]$$

and variance

$$\text{Var}(\kappa) \sim \frac{2(\bar{n}_H - 1)}{N\bar{n}_H(\bar{n}-1)^2} + \frac{(\bar{n}-\bar{n}_H)(1-4\bar{p}\,\bar{q})}{N\bar{n}\,\bar{n}_H(\bar{n}-1)^2\,\bar{p}\,\bar{q}} \quad . \qquad [16]$$

When the arithmetic and harmonic means of the number of judges are approximately equal, or when $\bar{p}$ is close to .5 (say, $.35 \leq p \leq .65$), then

$$\text{Var}(\kappa) \sim \frac{2(\bar{n}_H - 1)}{N\bar{n}_H(\bar{n}-1)^2} \quad . \qquad [17]$$

Finally, if the number of judges per subject is constant, say $n_i = n$ for $i = 1, \ldots, N$, then

$$\text{Var}(\kappa) \sim \frac{2}{Nn(n-1)} \quad , \qquad [18]$$

a result presented by Fleiss, Nee, and Landis (in press).

These variance formulas require the independence of judgments on one subject from those on another and are, therefore, not strictly correct if one or more of the judges are responsible for judging two or more of the subjects. The error seems to be small, however, provided that the total number of different judges is at least twice $\bar{n}$, the mean number of judges per subject.

The hypothesis that the underlying value of kappa is zero may be tested by referring the critical ratio

$$z = \frac{\kappa + \dfrac{1}{N(\bar{n}-1)}}{\sqrt{\text{Var}(\kappa)}} \qquad [19]$$

to tables of the standard normal distribution. This hypothesis is, admittedly, usually less interesting than questions appropriate to the non-null case when kappa is not hypothesized to be zero. One such question is whether two or more values of kappa (obtained, perhaps, after different intensities of training) are significantly different. Another concerns testing whether or not the underlying value of kappa is equal to some prespecified nonzero constant, or constructing a confidence interval for the parameter.

For such inferences, the variance appropriate to the non-null case should be used. Simple expressions for it do not exist, but Landis and Koch (1977b) refer to a computer program for its estimation.

Numerical Example

Consider the hypothetical data in Table 1. The number of subjects is $N = 15$, the mean number of judges per subject is $\bar{n} = 47/15 = 3.133$, and the overall proportion of positive judgments is $\bar{p} = 32/47 = 0.681$. The value of kappa is

$$\kappa = 1 - \frac{5.050}{15(2.133)(0.681)(0.319)} = 0.274. \qquad [20]$$

The expected value of kappa under the hypothesis of chance agreement is, by Equation 15, equal to −.031.

Table 1
Hypothetical Data on the Degree of Agreement on a Positive-Negative Trait Among Several Judges for 15 Subjects

Subject (i)	Number of Judges (n_i)	Number of Positives (x_i)
1	2	2
2	2	0
3	3	2
4	4	3
5	3	1
6	4	1
7	2	2
8	4	4
9	3	0
10	3	3
11	3	2
12	5	4
13	2	2
14	4	3
15	3	3
Total	47	32

Coding successes 1 and failures 0, the mean squares in the resulting analysis of variance table are

$$BMS = \frac{1}{14}(5.163) = 0.369 \qquad [21]$$

and

$$WMS = \frac{1}{15(2.133)}(5.050) = 0.158. \qquad [22]$$

The value of the constant n_o (Equation 7) is

$$n_o = 3.133 - \frac{0.838}{47} = 3.115, \qquad [23]$$

so the value of Landis and Koch's intraclass correlation coefficient of reliability (Equation 6) is

$$r = \frac{0.369 - 0.158}{0.369 + (2.115)(0.158)} = 0.300, \qquad [24]$$

slightly larger than the value of kappa. If BMS is redefined to have 15 instead of 14 as its divisor, it becomes equal to .344 and the value of r becomes .274, identical to the value of kappa.

The harmonic mean of the number of trials is

$$\bar{n}_H = \frac{15}{5.200} = 2.885, \qquad [25]$$

and the approximate variance of kappa (Equation 16) is equal to .0193. The statistical significance of kappa may be tested by referring the value of

$$z = \frac{\kappa - E(\kappa)}{\sqrt{Var(\kappa)}} = \frac{0.274+0.031}{0.14} = 2.18 \qquad [26]$$

to the standard normal distribution. Kappa is seen to be significantly different from 0 at the .05 level, although the degree of reliability is weak. If Equation 17 were used to find the variance, a value of .0191 would be obtained, and the resulting value of z would be 2.18, the same as above.

References

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

Cuzick, J. *Asymptotic normality of χ^2 in mxn tables with n large and small cell expectations.* Unpublished manuscript, 1977. Available from SIAM Institute for Mathematics and Society Study of Environmental Factors and Health.

Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 1971, *76*, 378–382.

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

Fleiss, J.L., & Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 1973, *33*, 613–619.

Fleiss, J.L., Nee, J., & Landis, J.R. The large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 1979, *86*, 974–977.

Haldane, J.B.S. The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small. *Biometrika*, 1939, *31*, 346–355.

Landis, J.R., & Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics*, 1977, *33*, 159–174. (a)

Landis, J.R., & Koch, G.G. A one-way components of variance model for categorical data. *Biometrics*, 1977, *33*, 671–679. (b)

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

Snedecor, G.W., & Cochran, W.G. *Statistical Methods* (6th Ed.). Ames: Iowa State University Press, 1967.

Acknowledgments

This work was supported in part by Grant MH 28655 from the National Institute of Mental Health and in part by a grant from the SIAM Institute for Mathematics and Society Study of Environmental Factors and Health.

Author's Address

Send requests for reprints or further information to Joseph L. Fleiss, Professor and Head, Division of Biostatistics, Columbia University School of Public Health, 600 West 168 Street, New York, NY 10032.