EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1973, 33, 613–619.

THE EQUIVALENCE OF WEIGHTED KAPPA AND THE INTRACLASS CORRELATION COEFFICIENT AS MEASURES OF RELIABILITY[1]

JOSEPH L. FLEISS
Biometrics Research, New York State Department of Mental Hygiene and Columbia University

JACOB COHEN
Department of Psychology, New York University

AN obvious factor restricting the comparability of categorical (nominal and ordinal) and quantitative (interval and ratio) data is that their descriptive statistics differ. For appraising reliability, for example, a useful measure of inter-rater agreement for categorical scales is provided by kappa (Cohen, 1960) or weighted kappa (Spitzer, Cohen, Fleiss and Endicott, 1967; Cohen, 1968a).

Kappa is the proportion of agreement corrected for chance, and scaled to vary from −1 to +1 so that a negative value indicates poorer than chance agreement, zero indicates exactly chance agreement, and a positive value indicates better than chance agreement. A value of unity indicates perfect agreement. The use of kappa implicitly assumes that all disagreements are equally serious. When the investigator can specify the relative seriousness of each kind of disagreement, he may employ weighted kappa, the proportion of weighted agreement corrected for chance.

For measuring the reliability of quantitative scales, the product-moment and intraclass correlation coefficients are widely

[1] This work was supported in part by Public Health Service Grants MH 08534 and MH 23964 from the National Institute of Mental Health. The authors are indebted to Dr. Robert L. Spitzer of Biometrics Research for suggesting the problem leading to this report.

used. Correspondences have been established between weighted kappa and the product-moment coefficient under restricted conditions (Cohen, 1960; 1968a). This paper establishes the equivalence of weighted kappa with the intraclass correlation coefficient under general conditions. Krippendorff (1970) demonstrated essentially the same result.

Weighted Kappa

Suppose that rater A distributes a sample of n subjects across the m categories of a categorical scale, and suppose that rater B independently does the same. Let n_{ij} denote the number of subjects assigned to category i by rater A and to category j by rater B; let $n_{i\cdot}$ denote the total number of subjects assigned to category i by rater A; and let $n_{\cdot j}$ denote the total number of subjects assigned to category j by rater B. Finally, let v_{ij} denote the disagreement weight associated with categories i and j. Typically, $v_{ii} = 0$, reflecting no disagreement when a subject is assigned to category i by both raters; and $v_{ij} > 0$ for $i \neq j$, reflecting some degree of disagreement when a subject is assigned to different categories by the two raters.

The mean observed degree of disagreement is

$$\bar{D}_o = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{m} n_{ij} v_{ij}, \tag{1}$$

and the mean degree of disagreement expected by chance (i.e., expected if A and B assign patients randomly in accordance with their respective base rates) is

$$\bar{D}_e = \frac{1}{n^2} \sum_{i=1}^{m} \sum_{j=1}^{m} n_{i\cdot} n_{\cdot j} v_{ij}. \tag{2}$$

Weighted kappa is then defined by

$$\kappa_w = \frac{\bar{D}_e - \bar{D}_o}{\bar{D}_e}. \tag{3}$$

Kappa is a special case of weighted kappa when $v_{ij} = 1$ for all $i \neq j$.

For convenience, weighted kappa has been defined here in terms of mean disagreement weights. Because weighted kappa is invariant under linear transformations of the weights (Cohen, 1968a), it is also interpretable as the proportion of weighted agreement corrected for chance.

The Problem of Interpretation

Given a sample value of weighted kappa, one may test for its statistical significance if the sample size n is large by referring the ratio of weighted kappa to its standard error to tables of the normal distribution. Exact small sample standard errors are given by Everitt (1968), and approximate large sample standard errors by Fleiss, Cohen and Everitt (1969). Assuming that the value of weighted kappa is significantly greater than zero, and even given that it is a proportion of *agreement*, there remains the problem: can its magnitude be compared with that obtained from a measure such as the intraclass correlation coefficient which is used with quantitive data and which is interpretable as a proportion of *variance*. An affirmative answer would provide a useful bridge over the gap between these two different levels of measurement.

Cohen has pointed out how, under certain conditions, kappa and weighted kappa may be interpreted as product-moment correlation coefficients. Specifically, for a 2×2 table whose marginal distributions are the same, kappa is precisely equal to the phi coefficient (Cohen, 1960). For a general $m \times m$ table with identical marginal distributions ($n_{i.} = n_{.i}$, $i = 1, \ldots, m$) and disagreement weights $v_{ij} = (i - j)^2$, weighted kappa is precisely equal to the product-moment correlation coefficient one would obtain if the nominal categories were scaled so that the first category was scored 1, the second category 2, etc. (Cohen, 1968a). Such a scaling is of course valid only when the categories may be ordered.

This paper establishes a more general property of weighted kappa. Specifically, if $v_{ij} = (i - j)^2$, and if the categories are scaled as above, then, irrespective of the marginal distributions, weighted kappa is identical with the intraclass correlation coefficient in which the mean difference between the raters is included as a component of variability.

Weighted Kappa in the Context of Two-Way Analysis of Variance

Let the ratings on the n subjects be quantified as described above. Define X_{k1} to be the ordinal number (either 1, 2, . . ., or m) of the category to which subject k was assigned by rater A,

and X_{k2} that of the category to which he was assigned by rater B. With $v_{ij} = (i - j)^2$ it may be shown (see equation 1) that

$$\bar{D}_o = \frac{1}{n} \sum_{k=1}^{n} (X_{k1} - X_{k2})^2, \tag{4}$$

and (see equation 2) that

$$\begin{aligned}\bar{D}_c &= \frac{1}{n^2} \sum_{i=1}^{m} \sum_{j=1}^{m} n_{i.} n_{.j} (i - j)^2 \\ &= \frac{1}{n^2} \sum_{i=1}^{m} n_{i.} i^2 \sum_{j=1}^{m} n_{.j} + \frac{1}{n^2} \sum_{i=1}^{m} n_{i.} \sum_{j=1}^{m} n_{.j} j^2 - \frac{2}{n^2} \sum_{i=1}^{m} n_{.i} i \sum_{j=1}^{m} n_{.j} j \\ &= \frac{1}{n} \sum_{k=1}^{n} X_{k1}^2 + \frac{1}{n} \sum_{k=1}^{n} X_{k2}^2 - 2\bar{X}_1 \bar{X}_2. \end{aligned} \tag{5}$$

In (5), $\bar{X}_1$ is the mean numerical rating given by rater A and $\bar{X}_2$ is the mean given by rater B.

Letting SS_r denote the sum of squares for raters in the analysis of variance of the X's, SS_s the sum of squares for subjects, and SS_e the error (residual) sum of squares, it is fairly easily checked that

$$\bar{D}_o = \frac{2}{n} (SS_r + SS_e) \tag{6}$$

and

$$\bar{D}_c = \frac{1}{n} (SS_s + 2SS_r + SS_e). \tag{7}$$

Thus,

$$\kappa_w = \frac{SS_s - SS_e}{SS_s + 2SS_r + SS_e}. \tag{8}$$

Suppose, now, that the n subjects are a random sample from a universe of subjects with variance σ_s^2, that the two raters are considered a random sample from a universe of raters with variance σ_r^2, and that the ratings are subject to a squared standard error of measurement σ_e^2. The sums of squares of the analysis of variance then estimate (see Scheffé, 1959, chapter 7)

$$E(SS_r) = \sigma_e^2 + n\sigma_r^2, \tag{9}$$

$$E(SS_s) = (n - 1)\sigma_e^2 + 2(n - 1)\sigma_s^2, \tag{10}$$

and

$$E(SS_e) = (n - 1)\sigma_e^2. \tag{11}$$

Thus, the numerator of (8) estimates

$$E(SS_s - SS_e) = 2(n - 1)\sigma_s^2, \tag{12}$$

and the denominator of (8) estimates

$$E(SS_s + 2SS_r + SS_e) = 2(n - 1)(\sigma_s^2 + \sigma_r^2 + \sigma_e^2) + 2(\sigma_r^2 + \sigma_e^2). \tag{13}$$

Therefore, (8) estimates, although not unbiasedly, the quantity

$$\rho' = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_e^2 + \frac{1}{n-1}(\sigma_r^2 + \sigma_e^2)}. \tag{14}$$

If n, the number of subjects, is at all large, then (8) in effect estimates

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_e^2}. \tag{15}$$

But ρ is the intraclass correlation coefficient between the ratings given a randomly selected subject by the randomly selected raters, for the covariance between two such ratings is σ_s^2, whereas the variance of any single such radomly generated rating is $\sigma_s^2 + \sigma_r^2 + \sigma_e^2$ (see Burdock, Fleiss and Hardesty, 1963). Thus κ_w is interpretable (aside from a term which goes to zero as n becomes large) as the intraclass correlation coefficient of reliability when systematic variability between raters is included as a component of total variation.

Comments

The above development was in terms of the measurement of agreement only on ordinal scales, and only when disagreement weights were taken as squared differences. The squaring of differences was admittedly arbitrary, but the scaling of errors by means of their squares is so common (see, e.g., Lehmann, 1959) that this convention requires little justification.

Of greater importance is the generalization to nominal scales of the interpretation of weighted kappa as an intraclass correlation coefficient. Such an interpretation will, it seems, be more or less valid provided that the disagreement weight v for two

categories increases more rapidly than the qualitative difference between them.

The latter idea provides a perspective from which the intraclass correlation coefficient may be viewed as a special case of weighted kappa. If the v_{ij}'s are viewed as squared distances between the categories of a nominal scale, they implicitly define a space of up to m-1 dimensions (Shepard, 1962). An m-point equal interval scale, on the other hand, explicitly defines a one-dimensional array with squared distances equal to $(i - j)^2$. As shown above, weighted kappa applied to the latter case necessarily equals its intraclass correlation. Thus the intraclass correlation coefficient is the special case of weighted kappa when the categories are equally spaced points along one dimension.

This perspective also makes it clear that in assigning weights v_{ij} in the nominal case one is in effect quantifying a nominal scale of m categories, as in multidimensional scaling (Shepard, 1962) or in multiple regression analysis with nominal scale coding (Cohen, 1968b), by going beyond one dimension to as many as m-1 dimensions.

Application

A multidimensional structure for nominal scales was implicitly incorporated into a scaling of differences in psychiatric diagnosis (Spitzer, Cohen, Fleiss and Endicott, 1967). Two diagnostic categories which differed markedly in terms of severity or prognosis were given a disagreement weight appreciably in excess of the weight associated with two similar categories. Values of weighted kappa in the interval .4 to .6 have typically been found for agreement on psychiatric diagnosis (Spitzer and Endicott, 1968; 1969).

Some widely used numerical scales of psychopathology, on the other hand, typically have reliabilities in the interval .7 to .9 when reliability is measured by the intraclass correlation coefficient (Spitzer, Fleiss, Endicott and Cohen, 1967). Given the correspondence established above between weighted kappa and the intraclass correlation coefficient, and given the rationale used in scaling diagnostic disagreement, it seems possible to affirm that agreement on psychiatric diagnosis is poorer than agreement on numerical descriptions of psychopathology.

REFERENCES

Burdock, E. I., Fleiss, J. L., and Hardesty, A. S. A new view of inter-observer agreement. *Personnel Psychology,* 1963, 16, 373–384.

Cohen, J. A coefficient of agreement for nominal scales. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1960, 20, 37–46.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit: *Psychological Bulletin,* 1968, 70, 213–220. (a)

Cohen, J. Multiple regression as a general data-analytic system. *Psychological Bulletin,* 1968, 70, 426–443. (b)

Everitt, B. S. Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology,* 1968, 21, 97–103.

Fleiss, J. L., Cohen, J. and Everitt, B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin,* 1969, 72, 323–327.

Krippendorff, K. Bivariate agreement coefficients for reliability of data. In E. F. Borgatta and G. W. Bohrnstedt (Eds.) *Social Methodology 1970.* San Francisco: Jossey-Bass, 1970.

Lehmann, E. L. *Testing statistical hypotheses.* New York: Wiley, 1959.

Scheffé, H. *The analysis of variance.* New York: Wiley, 1959.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika,* 1962, 27, 125–140.

Spitzer, R. L., Cohen, J., Fleiss, J. L. and Endicott, J. Quantification of agreement in psychiatric diagnosis. *Archives of General Psychiatry,* 1967, 17, 83–87.

Spitzer, R. L. and Endicott, J. Diagno: A computer program for psychiatric diagnosis utilizing the differential diagnostic procedure. *Archives of General Psychiatry,* 1968, 18, 746–756.

Spitzer, R. L. and Endicott, J. Diagno II: Further developments in a computer program for psychiatric diagnosis. *American Journal of Psychiatry,* 1969, 125 (Jan. supp.), 12–21.

Spitzer, R. L., Fleiss, J. L., Endicott, J. and Cohen, J. Mental status schedule: Properties of factor-analytically derived scales. *Archives of General Psychiatry,* 1967, 16, 479–493.