Psychological Bulletin
1969, Vol. 72, No. 5, 323–327

LARGE SAMPLE STANDARD ERRORS OF KAPPA AND WEIGHTED KAPPA[1]

JOSEPH L. FLEISS[2]
Biometrics Research, New York State Department of Mental Hygiene

JACOB COHEN
New York University

AND B. S. EVERITT
Institute of Psychiatry, University of London

Two statistics, kappa and weighted kappa, are available for measuring agreement between two raters on a nominal scale. Formulas for the standard errors of these two statistics have been given in the literature, but they are in error. The errors seem to be in the direction of overestimation, so that the use of the incorrect formulas results in conservative significance tests and confidence intervals. Valid formulas for the approximate large-sample variances are given, and their calculation is illustrated using a numerical example.

The statistics kappa (Cohen, 1960) and weighted kappa (Cohen, 1968) were introduced to provide coefficients of agreement between two raters for nominal scales. Kappa is appropriate when all disagreements may be considered equally serious, and weighted kappa is appropriate when the relative seriousness of the different possible disagreements can be specified.

The papers describing these two statistics also present expressions for their standard errors. These expressions are incorrect, having been derived from the contradictory assumptions of fixed marginal totals and binomial variation of cell frequencies. Everitt (1968) derived the exact variances of weighted and unweighted kappa when the parameters are zero by assuming a generalized hypergeometric distribution. He found these expressions to be far too complicated for routine use, and offered, as alternatives, expressions derived by assuming binomial distributions. These alternative expressions are incorrect, essentially for the same reason as above.

Assume that N subjects are distributed into k^2 cells by each of them being assigned to one of k categories by one rater and, independently, to one of the same k categories by a second rater. Let p_{ij} be the proportion of subjects placed in the i, jth cell; let

$$p_{i.} = \sum_{j=1}^{k} p_{ij}, \qquad [1]$$

the proportion of subjects placed in the ith row; let

$$p_{.j} = \sum_{i=1}^{k} p_{ij}, \qquad [2]$$

the proportion of subjects placed in the jth column; and let w_{ij}, assumed without loss of generality to lie between 0 and 1, be the weight assigned to the i, jth cell. Then, with

$$p_o = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij} \qquad [3]$$

and

$$p_c = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i.} p_{.j}, \qquad [4]$$

weighted kappa is defined by

$$\hat{\kappa}_w = \frac{p_o - p_c}{1 - p_c}. \qquad [5]$$

Define

$$\bar{w}_{i.} = \sum_{j=1}^{k} w_{ij} p_{.j}, \qquad [6]$$

a weighted average of the weights in the ith row, and

$$\bar{w}_{.j} = \sum_{i=1}^{k} w_{ij} p_{i.}, \qquad [7]$$

[1] This work was supported in part by Grant MH 08534 and in part by Grant MH 09191, both from the National Institute of Mental Health.

[2] Also at Columbia University.

Requests for reprints should be addressed to Joseph L. Fleiss, Biometrics Research, 722 West 168th Street, New York, New York 10032.

a weighted average of the weights in the jth column. The estimated large sample variance of $\hat{\kappa}_w$, useful in setting confidence limits or in comparing two independent values of $\hat{\kappa}_w$, is

$$\widehat{\text{Var}}(\hat{\kappa}_w)=\frac{1}{N(1-p_c)^4}\{\sum_{i=1}^{k}\sum_{j=1}^{k}p_{ij}[w_{ij}(1-p_c)-(\bar{w}_{i.}+\bar{w}_{.j})(1-p_o)]^2-(p_op_c-2p_c+p_o)^2\}. \qquad [8]$$

The variance was derived by using the classic result appearing, for example, in Rao (1965, p. 321), and by imposing no restrictions on the observed array other than fixing N. In particular, the validity of Equation 8 does not require fixed marginals.

The estimated variance of $\hat{\kappa}_w$ when there is no association between the two raters' assignments (a sufficient but not necessary condition for the population value of weighted kappa, κ_w, to be zero) is found by replacing p_o by p_c and p_{ij} by $p_{i.}p_{.j}$ in Equation 8:

$$\widehat{\text{Var}}_0(\hat{\kappa}_w)=\frac{1}{N(1-p_c)^2}\{\sum_{i=1}^{k}\sum_{j=1}^{k}p_{i.}p_{.j}\times[w_{ij}-(\bar{w}_{i.}+\bar{w}_{.j})]^2-p_c^2\}. \qquad [9]$$

Expression 9 may be used in testing the hypothesis that $\kappa_w = 0$.

Estimated large sample variances of unweighted kappa, $\hat{\kappa}$, follow from Expressions 8 and 9 by noting that $\hat{\kappa}$ is a special case of $\hat{\kappa}_w$ with $w_{ij} = 1$ for $i = j$ and $w_{ij} = 0$ for $i \neq j$. Thus, with

$$p_o = \sum_{i=1}^{k} p_{ii} \qquad [10]$$

and

$$p_c = \sum_{i=1}^{k} p_{i.}p_{.i}, \qquad [11]$$

$\hat{\kappa}$ is given by

$$\hat{\kappa} = \frac{p_o - p_c}{1 - p_c}. \qquad [12]$$

TABLE 1

Hypothetical Data on 200 Cases to Illustrate the Computation of the Variances of Weighted Kappa

Rater B	Statistic	Rater A			$p_{i.}$	$\bar{w}_{i.}$
		1	2	3		
1	a	1	0	.4444		.6944
	b	.53	.05	.02	.60	
	c	.39	.15	.06		
	d	1.3388	1.0611	1.2611		
	e	.021855	.051082	.005805		
	f	.114785	1.125933	.666999		
2	a	0	1	.6667		.3167
	b	.11	.14	.05	.30	
	c	.195	.075	.03		
	d	.9611	.6834	.8834		
	e	.041908	.082619	.010104		
	f	.923713	.100236	.046959		
3	a	.4444	.6667	1		.5556
	b	.01	.06	.03	.10	
	c	.065	.025	.01		
	d	1.2000	.9223	1.1223		
	e	.003991	.008507	.037617		
	f	.570931	.065331	.014957		
	$p_{.j}$	.65	.25	.10	1.00	
	$\bar{w}_{.j}$	.64444	.3667	.5667		

Note.—For each cell, the six entries are as follows: $a = w_{ij}$; $b = p_{ij}$; $c = p_{i.}p_{.j}$; $d = \bar{w}_{i.}+\bar{w}_{.j}$; $e = [w_{ij}(1-p_o)-(\bar{w}_{i.}+\bar{w}_{.j})(1-p_o)]^2$; $f = [w_{ij}-(\bar{w}_{i.}+\bar{w}_{.j})]^2$.

The estimated large sample variance of $\hat{\kappa}$ is (note that $\bar{w}_{i.}$ becomes $p_{.i}$ and that $\bar{w}_{.j}$ becomes $p_{j.}$)

$$\widehat{\mathrm{Var}}(\hat{\kappa})=\frac{1}{N(1-p_c)^4}\{\sum_{i=1}^{k} p_{ii} \times[(1-p_c)-(p_{.i}+p_{i.})(1-p_o)]^2 + (1-p_o)^2 \sum_{i=1}^{k}\sum_{\substack{j=1\\ i\neq j}}^{k} p_{ij}(p_{.i}+p_{j.})^2 - (p_o p_c - 2p_c + p_o)^2\}, \quad [13]$$

and the variance appropriate to the case of no association is

$$\widehat{\mathrm{Var}}_0(\hat{\kappa})=\frac{1}{N(1-p_c)^2}\{\sum_{i=1}^{k} p_{i.}p_{.i} \times[1-(p_{.i}+p_{i.})]^2 + \sum_{i=1}^{k}\sum_{\substack{j=1\\ i\neq j}}^{k} p_{i.}p_{.j}(p_{.i}+p_{j.})^2 - p_c^2\}. \quad [14]$$

The study of many numerical examples indicates that the variance expressions given by Cohen (1960, 1968), and the nonexact formulas given by Everitt (1968) overestimate the variance. Thus, their use results in conservative significance tests and confidence intervals.

The hypothetical data in Table 1 are used to illustrate the calculation of the variances of weighted kappa. The *a* entry in each cell is the weight, w_{ij}. The *b* entry is the observed proportion out of $N = 200$, p_{ij}. The *c* entry is the proportion expected by chance, $p_{i.}p_{.j}$. After bordering the table with the average weights $\bar{w}_{i.}$ (Expression 6) and $\bar{w}_{.j}$ (Expression 7), the *d* entries in the cells, $\bar{w}_{i.} + \bar{w}_{.j}$, may be calculated.

The observed weighted proportion of agreement is obtained by multiplying the *a* entry in each cell by the *b* entry and summing over all cells; it is

$$p_o = .787. \quad [15]$$

The weighted proportion of agreement expected by chance is obtained by multiplying the *a* and *c* entries and summing over all cells; it is

$$p_c = .567. \quad [16]$$

Thus,

$$\hat{\kappa}_w = \frac{.787 - .567}{1 - .567} = .508. \quad [17]$$

The *e* entry in each cell is the square of the quantity: $(1 - p_c)$ times the *a* entry minus $(1 - p_o)$ times the *d* entry, or, $[w_{ij}(1 - p_c) - (\bar{w}_{i.} + \bar{w}_{.j})(1 - p_o)]^2$. Multiplying the *b* and *e* entries and summing over all cells yields

$$\sum_{i=1}^{3}\sum_{j=1}^{3} p_{ij}[w_{ij}(1 - p_c) - (\bar{w}_{i.}+\bar{w}_{.j})(1-p_o)]^2 = .032614. \quad [18]$$

Since

$$(p_o p_c - 2p_c + p_o)^2 = (.787\times.567 - 2\times.567 + .787)^2 = .009846 \quad [19]$$

and since

$$(1 - p_c)^4 = (.433)^4 = .035152, \quad [20]$$

therefore, by Equation 8

$$\widehat{\mathrm{Var}}(\hat{\kappa}_w) = \frac{.032614 - .009846}{200\times.035152} = .003239. \quad [21]$$

The formula derived by Fleiss and given by Cohen (1968, Equation 10) yields, with Cohen's $v_{ij} = 1 - w_{ij}$, an estimated variance of .003630, which is seen to be larger than the value in Equation 21.

The *f* entry in each cell is the square of the difference between the *a* and *d* entries, or $[w_{ij} - (\bar{w}_{i.} + \bar{w}_{.j})]^2$. Multiplying the *f* and the *c* entries and summing over all cells yields

$$\sum_{i=1}^{3}\sum_{j=1}^{3} p_{i.}p_{.j}[w_{ij} - (\bar{w}_{i.}+\bar{w}_{.j})]^2 = .481620. \quad [22]$$

Since

$$p_c^2 = .321489 \quad [23]$$

and since

$$(1 - p_c)^2 = .187489, \quad [24]$$

therefore, by Equation 9, the estimated variance under the hypothesis that $\kappa_w = 0$ is

$$\widehat{\mathrm{Var}}_0(\hat{\kappa}_w) = \frac{.481620 - .321489}{200\times.187489} = .004270. \quad [25]$$

The formula given by Cohen (1968, Equation 13) yields an estimated variance of .005403, which is seen to be larger than the value in

TABLE 2

Hypothetical Data on 200 Cases to Illustrate the Computation of the Variances of Kappa

Rater B	Statistic	Rater A			$p_{i.}$	$\bar{w}_{i.} = p_{.i}$
		1	2	3		
1	a	.53	.05	.02	.60	.65
	b	.39	.15	.06		
	c	1.25	.95	.75		
	d	.022500	—	—		
	e	—	.9025	.5625		
	f	.0625	—	—		
2	a	.11	.14	.05	.30	.25
	b	.195	.075	.03		
	c	.85	.55	.35		
	d	—	.129600	—		
	e	.7225	—	.1225		
	f	—	.2025	—		
3	a	.01	.06	.03	.10	.10
	b	.065	.025	.01		
	c	.70	.40	.20		
	d	—	—	.216225		
	e	.4900	.1600	—		
	f	—	—	.6400		
	$p_{.j}$	.65	.25	.10	1.00	
	$\bar{w}_{.j} = p_{j.}$	.60	.30	.10		

Note.—For each cell, the six entries are as follows: $a = p_{ij}$; $b = p_{i.}p_{.j}$; $c = p_{.i}+p_{j.}$; $d = [(1-p_c)-(p_{.i}+p_{i.})(1-p_o)]^2$; $e = (p_{.i}+p_{j.})^2$; $= [1-(p_{.i}+p_{i.})]^2$.

Equation 25. Everitt (1968, p. 102) found the exact variance to be .004417, which would indicate that the expression given in Equation 9 somewhat underestimates the exact value.

Table 2 illustrates the calculation of the variances of unweighted kappa, using the same hypothetical p_{ij}'s as in Table 1, here the a entries in the cells. The b entry is the proportion expected by chance, $p_{i.}p_{.j}$. After bordering the table with the column of $\bar{w}_{i.}$'s—which are now simply the column proportions, $\{p_{.i}\}$—and with the row of $\bar{w}_{.j}$'s—which are now the row proportions, $\{p_{j.}\}$—the c entries in the cells, $\bar{w}_{i.} + \bar{w}_{.j} = p_{.i} + p_{j.}$, are calculated.

The observed proportion of agreement (Equation 10) is obtained by summing the a entries for the agreement cells (those with $i = j$) only; it is

$$p_o = .70. \qquad [26]$$

The chance proportion of agreement (Equation 11) is obtained by summing the b entries for the agreement cells; it is

$$p_c = .475. \qquad [27]$$

$\hat{\kappa}$ (Equation 12) is then

$$\hat{\kappa} = \frac{.70 - .475}{1 - .475} = .429. \qquad [28]$$

The d entry is found only for the cells with $i = j$. It is the square of the quantity: $(1 - p_c)$ minus $(1-p_o)$ times the c entry, or $[(1-p_c) - (p_{.i}+p_{i.})(1-p_o)]^2$. Multiplication of the d and the a entries and summing for the diagonal cells yields

$$\sum_{i=1}^{3} p_{ii}[(1 - p_c) - (p_{.i} + p_{i.})(1 - p_o)]^2 = .036556. \qquad [29]$$

The e entry is found only for the cells with $i \neq j$. It is simply the square of the c entry, $(p_{.i}+p_{j.})^2$. Multiplication of the e and the a entries and summing for the off-diagonal cells yields

$$\sum_{i=1}^{3}\sum_{\substack{j=1\\i\neq j}}^{3} p_{ij}(p_{.i}+p_{j.})^2 = .156475. \quad [30]$$

Multiplication of this last value by $(1-p_o)^2 = .09$ yields

$$(1-p_o)^2\sum_{i=1}^{3}\sum_{\substack{j=1\\i\neq j}}^{3} p_{ij}(p_{.i}+p_{j.})^2 = .014083. \quad [31]$$

Since

$$(p_op_c-2p_c+p_o)^2 = (.70\times.475-2\times.475+.70)^2 = .006806 \quad [32]$$

and since

$$(1-p_c)^4 = (.525)^4 = .075969, \quad [33]$$

therefore, by Equation 13

$$\widehat{\mathrm{Var}}(\hat{\kappa}) = \frac{.036556+.014083-.006806}{200\times.075969} = .002885. \quad [34]$$

The formula given by Cohen (1960, Equation 7 and repeated in 1968 as Equation 2) yields an estimated variance of .003810, which is seen to be larger than the value in Equation 34.

The f entry is found only for the agreement cells, that is, those with $i=j$. It is the square of the difference between unity and the c entry, or $[1-(p_{.i}+p_{i.})]^2$. Multiplication of the f and the b entries and summing for the diagonal cells yields

$$\sum_{i=1}^{3} p_{i.}p_{.i}[1-(p_{.i}+p_{i.})]^2 = .045962. \quad [35]$$

For the disagreement cells, that is, those with $i \neq j$, summing the products of the b and e entries yields

$$\sum_{i=1}^{3}\sum_{\substack{j=1\\i\neq j}}^{3} p_{i.}p_{.j}(p_{.i}+p_{j.})^2 = .349538. \quad [36]$$

Since

$$p_c^2 = (.475)^2 = .225625 \quad [37]$$

and since

$$(1-p_c)^2 = (.525)^2 = .275625, \quad [38]$$

therefore, by Equation 14, the estimated variance under the hypothesis that $\kappa = 0$ is

$$\widehat{\mathrm{Var}}_0(\hat{\kappa}) = \frac{.045962+.349538-.225625}{200\times.275625} = .003082. \quad [39]$$

The formula given by Cohen (1960, Equation 10 and repeated in 1968 as Equation 3) yields an estimated variance under the hypothesis that $\kappa = 0$ of .004524, which is again seen to be larger than the value in Expression 39. The exact formula of Everitt (1968, p. 102) gives .0031, in agreement with the value in Expression 39.

For reference purposes, the illustrative example in Cohen (1960, Table 2; 1968, Table 1) gives the following correct variances with the previous results in parentheses:

$$\widehat{\mathrm{Var}}(\hat{\kappa}_w) = .005707\ (.008118);$$
$$\widehat{\mathrm{Var}}_0(\hat{\kappa}_w) = .003570\ (.008391);$$
$$\widehat{\mathrm{Var}}(\hat{\kappa}) = .002601\ (.003016);$$

and

$$\widehat{\mathrm{Var}}_0(\hat{\kappa}) = .002702\ (.003475).$$

REFERENCES

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.

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

Rao, C. R. *Linear statistical inference and its applications.* New York: Wiley, 1965.

(Received January 13, 1969)