CHAPTER 18

The Measurement of Interrater Agreement

The statistical methods described in the preceding chapter for controlling for error are applicable only when the rates of misclassification are known from external sources or are estimable by applying a well-defined standard classification procedure to a subsample of the group under study. For some variables of importance, however, no such standard is readily apparent.

To assess the extent to which a given characterization of a subject is reliable, it is clear that we must have a number of subjects classified more than once, for example by more than one rater. The degree of agreement among the raters provides no more than an upper bound on the degree of accuracy present in the ratings, however. If agreement among the raters is good, then there is a possibility, but by no means a guarantee, that the ratings do in fact reflect the dimension they are purported to reflect. If their agreement is poor, on the other hand, then the usefulness of the ratings is severely limited, for it is meaningless to ask what is associated with the variable being rated when one cannot even trust those ratings to begin with.

In this chapter we consider the measurement of interrater agreement when the ratings are on categorical scales. Section 18.1 is devoted to the case of the same two raters per subject. Section 18.2 considers weighted kappa to incorporate a notion of distance between rating categories. Section 18.3 is devoted to the case of multiple ratings per subject with different sets of raters. Applications to other problems are indicated in Section 18.4. Section 18.5* relates the results of the preceding sections to the theory presented in Chapter 15 on correlated binary variables.

Statistical Methods for Rates and Proportions, Third Edition
By Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik
ISBN 0-471-52629-0

Table 18.1. *Diagnoses on n = 100 subjects by two raters*

	Rater *B*			
Rater *A*	Psychotic	Neurotic	Organic	Total
Psychotic	0.75	0.01	0.04	0.80
Neurotic	0.05	0.04	0.01	0.10
Organic	0	0	0.10	0.10
Total	0.80	0.05	0.15	1.00

18.1. THE SAME PAIR OF RATERS PER SUBJECT

Suppose that each of a sample of n subjects is rated independently by the same two raters, with the ratings being on a categorical scale consisting of k categories. Consider the hypothetical example of Table 18.1, in which each cell entry is the proportion of all subjects classified into one of $k = 3$ diagnostic categories by rater A and into another by rater B. Thus, for example, 5% of all subjects were diagnosed neurotic by rater A and psychotic by rater B.

Suppose it is desired to measure the degree of agreement on each category separately as well as across all categories. The analysis begins by collapsing the original $k \times k$ table into a 2×2 table in which all categories other than the one of current interest are combined into a single "all others" category. Table 18.2 presents the results in general, as well as for neurosis from Table 18.1 in particular. It must be borne in mind that the entries a, b, c, and d in the general table refer to *proportions* of subjects, not to their numbers.

The simplest and most frequently used index of agreement is the overall proportion of agreement, say

$$p_o = a + d. \tag{18.1}$$

Table 18.2. *Data for measuring agreement on a single category*

General				For Neurosis			
	Rater *B*				Rater *B*		
Rater *A*	Given Category	All Others	Total	Rater *A*	Neurosis	All Others	Total
Given category	*a*	*b*	p_1	Neurosis	0.04	0.06	0.10
All others	*c*	*d*	q_1	All others	0.01	0.89	0.90
Total	p_2	q_2	1	Total	0.05	0.95	1.00

Table 18.3. *Values of several indices of agreement from data of Table 18.1*

Category	p_o	p_s	λ_r	p'_s	A	κ
Psychotic	0.90	0.94	0.88	0.75	0.84	0.69
Neurotic	0.93	0.53	0.06	0.96	0.75	0.50
Organic	0.95	0.80	0.60	0.97	0.89	0.77

p_o, or a simple variant of it such as $2p_o - 1$, has been proposed as the agreement index of choice by Holley and Guilford (1964) and by Maxwell (1977). For neurosis, the overall proportion of agreement is

$$p_o = 0.04 + 0.89 = 0.93.$$

This value, along with the overall proportions of agreement for the other two categories, is given in the column labeled p_o in Table 18.3. The conclusion that might be drawn from these values is that agreement is, effectively, equally good on all three categories, with agreement on organic disorders being somewhat better than on neurosis, and agreement on neurosis being somewhat better than on psychosis.

Suppose the category under study is rare, so that the proportion d, representing agreement on absence, is likely to be large and thus to inflate the value of p_o. A number of indices of agreement have been proposed that are based only on the proportions a, b, and c. Of all of them, only the so-called proportion of specific agreement, say

$$p_s = \frac{2a}{2a + b + c} = \frac{a}{\bar{p}}, \tag{18.2}$$

where $\bar{p} = (p_1 + p_2)/2$, has a sensible probabilistic interpretation. Let one of the two raters be selected at random, and let attention be focused on the subjects assigned to the category of interest. The quantity p_s is the conditional probability that the second rater will also make an assignment to that category, given that the randomly selected first rater did. This index was first proposed by Dice (1945) as a measure of similarity.

The proportion of specific agreement on neurosis is

$$p_s = \frac{2 \times 0.04}{2 \times 0.04 + 0.06 + 0.01} = 0.53,$$

and the values for all three categories are presented in the column headed p_s in Table 18.3. The conclusions based on p_s are rather different from those based on p_o. Agreement now seems best on psychosis, rather less good on organic disorders, and much poorer than either on neurosis.

Define $\bar{q} = 1 - \bar{p}$, or

$$\bar{q} = \frac{1}{2}(q_1 + q_2) = d + \frac{b+c}{2}, \tag{18.3}$$

and suppose that $\bar{q} > \bar{p}$. Goodman and Kruskal (1954) proposed

$$\lambda_r = \frac{(a+d) - \bar{q}}{1 - \bar{q}} = \frac{2a - (b+c)}{2a + (b+c)} \tag{18.4}$$

as an index of agreement; it is motivated less by notions of agreement than by a consideration of the frequencies of correct predictions of a subject's category when predictions are made with and without knowledge of the joint ratings. λ_r assumes its maximum value of $+1$ when there is complete agreement, but assumes its minimum value of -1 whenever $a = 0$, irrespective of the value of d [not, as Goodman and Kruskal (1954, p. 758) imply, only when $a + d = 0$].

For neurosis,

$$\lambda_r = \frac{2 \times 0.04 - (0.06 + 0.01)}{2 \times 0.04 + (0.06 + 0.01)} = 0.06,$$

and the values of λ_r for all three categories are listed under the indicated column of Table 18.3. Because of the identity

$$\lambda_r = 2p_s - 1, \tag{18.5}$$

the categories are ordered on λ_r exactly as on p_s.

The proportion of specific agreement ignores the proportion d. If, instead, we choose to ignore a, we would calculate the corresponding index, say

$$p_s' = \frac{d}{\bar{q}} = \frac{2d}{2d + b + c}, \tag{18.6}$$

where $\bar{q} = 1 - \bar{p}$. For neurosis

$$p_s' = \frac{2 \times 0.89}{2 \times 0.89 + 0.06 + 0.01} = 0.96,$$

and this value and the other two are presented in the indicated column of Table 18.3. Yet a different picture emerges from these values than from earlier ones. Agreement (with respect to absence) on organic disorders and on neurosis seems to be equally good and apparently substantially better than on psychosis.

Rather than having to choose between p_s and p'_s, Rogot and Goldberg (1966) proposed simply taking their mean, say

$$A = \frac{1}{2}(p_s + p'_s) = \frac{a}{p_1 + p_2} + \frac{d}{q_1 + q_2}, \tag{18.7}$$

as an index of agreement. For neurosis,

$$A = \frac{0.04}{0.10 + 0.05} + \frac{0.89}{0.90 + 0.95} = 0.75.$$

As seen in the indicated column of Table 18.3, the index A orders the three categories in yet a new way: agreement on organic disorders is better than on psychosis, and agreement on organic disorders and on psychosis is better than on neurosis.

Yet other indices of agreement between two raters have been proposed (e.g., Fleiss, 1965; Armitage, Blendis, and Smyllie, 1966; Rogot and Goldberg, 1966; and Bennett, 1972), but it should already be clear that there must be more to the measurement of interrater agreement than the arbitrary selection of an index of agreement.

The new dimension is provided by a realization that, except in the most extreme circumstances (either $p_1 = q_2 = 0$ or $p_2 = q_1 = 0$), some degree of agreement is to be expected by chance alone (see Table 18.4). For example, if rater A employs one set of criteria for distinguishing between the presence and the absence of a condition, and if rater B employs an entirely different and independent set of criteria, then *all* the observed agreement is explainable by chance.

Different opinions have been stated on the need to incorporate chance-expected agreement into the assessment of interrater reliability. Rogot and Goldberg (1966), for example, emphasize the importance of contrasting observed with expected agreement when comparisons are to be made between different pairs of raters or different kinds of subjects. Goodman and

Table 18.4. ***Chance-expected proportions of joint judgments by two raters, for data of Table 18.2***

General				For Neurosis			
	Rater B				Rater B		
Rater A	Given Category	All Others	Total	Rater A	Neurosis	All Others	Total
Given category	p_1p_2	p_1q_2	p_1	Neurosis	0.005	0.095	0.10
All others	q_1p_2	q_1q_2	q_1	All others	0.045	0.855	0.90
Total	p_2	q_2	1	Total	0.05	0.95	1

Kruskal (1954, p. 758), on the other hand, contend that chance-expected agreement need not cause much concern, that the observed degree of agreement may usually be assumed to be in excess of chance. (Even if one is willing to grant this assumption, one should nevertheless check whether the excess is trivially small or substantially large.)

Armitage, Blendis, and Smyllie (1966, p. 102) occupy a position between that of Rogot and Goldberg and that of Goodman and Kruskal. They appreciate the necessity for introducing chance-expected agreement whenever different sets of data are being compared, but claim that too much uncertainty exists as to how the correction for chance is to be incorporated into the measure of agreement.

There does exist, however, a natural means for correcting for chance. Consider any index that assumes the value 1 when there is complete agreement. Let I_o denote the observed value of the index (calculated from the proportions in Table 18.2), and let I_e denote the value expected on the basis of chance alone (calculated from the proportions in Table 18.4).

The obtained excess beyond chance is $I_o - I_e$, whereas the maximum possible excess is $1 - I_e$. The ratio of these two differences is called *kappa*,

$$\hat{\kappa} = \frac{I_o - I_e}{1 - I_e}. \tag{18.8}$$

Kappa is a measure of agreement with desirable properties. If there is complete agreement, $\hat{\kappa} = +1$. If observed agreement is greater than or equal to chance agreement, $\hat{\kappa} \geq 0$, and if observed agreement is less than or equal to chance agreement, $\hat{\kappa} \leq 0$. The minimum value of $\hat{\kappa}$ depends on the marginal proportions. If they are such that $I_e = 0.5$, then the minimum equals -1. Otherwise, the minimum is between -1 and 0.

It may be checked by simple algebra that, *for each of the indices of agreement defined above*, the same value of $\hat{\kappa}$ results after the chance-expected value is incorporated as in (18.8) (see Problem 18.1):

$$\hat{\kappa} = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1}. \tag{18.9}$$

An important unification of various approaches to the indexing of agreement is therefore achieved by introducing a correction for chance-expected agreement.

For neurosis,

$$\hat{\kappa} = \frac{2(0.04 \times 0.89 - 0.06 \times 0.01)}{0.10 \times 0.95 + 0.05 \times 0.90} = 0.50.$$

This value and the other two are presented in the final column of Table 18.3. They are close to those found by Spitzer and Fleiss (1974) in a review of the

literature on the reliability of psychiatric diagnosis. Agreement is best on organic disorders, less good on psychosis, and poorest on neurosis.

The kappa statistic was first proposed by Cohen (1960). Variants of kappa have been proposed by Scott (1955) and by Maxwell and Pilliner (1968). All have interpretations as *intraclass correlation coefficients* (see Ebel, 1951). The intraclass correlation coefficient is a widely used measure of interrater reliability for the case of quantitative ratings. As shown by Fleiss (1975) and Krippendorff (1970), only kappa is identical (except for a term involving the factor $1/n$, where n is the number of subjects) to that version of the intraclass correlation coefficient due to Bartko (1966) in which a difference between the raters in their base rates (i.e., a difference between p_1 and p_2) is considered a source of unwanted variability.

Landis and Koch (1977a) have characterized different ranges of values for kappa with respect to the degree of agreement they suggest. For most purposes, values greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values below 0.40 or so may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance.

Often, a composite measure of agreement across all categories is desired. An overall value of kappa may be defined as a weighted average of the individual kappa values, where the weights are the denominators of the individual kappas [i.e., the quantities $p_1q_2 + p_2q_1$ in (18.9)]. An equivalent and more suggestive formula is based on arraying the data as in Table 18.5.

The overall proportion of observed agreement is, say,

$$p_o = \sum_{i=1}^{k} p_{ii}, \tag{18.10}$$

and the overall proportion of chance-expected agreement is, say,

$$p_e = \sum_{i=1}^{k} p_{i.}p_{.i}. \tag{18.11}$$

Table 18.5. ***Joint proportions of ratings by two raters on a scale with k categories***

	Rater *B*				
Rater *A*	1	2	⋯	*k*	Total
1	p_{11}	p_{12}	⋯	p_{1k}	$p_{1.}$
2	p_{21}	p_{22}	⋯	p_{2k}	$p_{2.}$
⋮	⋮	⋮		⋮	⋮
k	p_{k1}	p_{k2}	⋯	p_{kk}	$p_{k.}$
Total	$p_{.1}$	$p_{.2}$	⋯	$p_{.k}$	1

The overall value of kappa is then, say,

$$\hat{\bar{\kappa}} = \frac{p_o - p_e}{1 - p_e}. \tag{18.12}$$

For the data of Table 18.1,

$$p_o = 0.75 + 0.04 + 0.10 = 0.89$$

and

$$p_e = 0.80 \times 0.80 + 0.10 \times 0.05 + 0.10 \times 0.15 = 0.66,$$

so that

$$\hat{\bar{\kappa}} = \frac{0.89 - 0.66}{1 - 0.66} = 0.68.$$

For testing the hypothesis that the ratings are independent (so that the underlying value of kappa is zero), Fleiss, Cohen, and Everitt (1969) showed that the appropriate standard error of kappa is estimated by

$$\widehat{\text{se}}_0(\hat{\bar{\kappa}}) = \frac{1}{(1 - p_e)\sqrt{n}} \sqrt{p_e + p_e^2 - \sum_{i=1}^{k} p_{i.} p_{.i} (p_{i.} + p_{.i})}, \tag{18.13}$$

where p_e is defined in (18.11). The hypothesis may be tested against the alternative that agreement is better than chance would predict by referring the quantity

$$z = \frac{\hat{\bar{k}}}{\widehat{\text{se}}_0(\hat{\bar{\kappa}})} \tag{18.14}$$

to tables of the standard normal distribution and rejecting the hypothesis if z is sufficiently large (a one-sided test is more appropriate here than a two-sided test).

For the data at hand,

$$\widehat{\text{se}}_0(\hat{\bar{\kappa}}) = \frac{1}{(1 - 0.66)\sqrt{100}} \sqrt{0.66 + 0.66^2 - 1.0285} = 0.076$$

and

$$z = \frac{0.68}{0.076} = 8.95.$$

The overall value of kappa is therefore statistically highly significant, and, by virtue of its magnitude, it indicates a good degree of agreement beyond chance.

Table 18.6. *Kappas for individual categories and across all categories of Table 18.1*

Category	p_o	p_e	$\hat{\kappa}$	$\widehat{\text{se}}_0(\hat{\kappa})$	z
Psychotic	0.90	0.68	0.69	0.100	6.90
Neurotic	0.93	0.86	0.50	0.093	5.38
Organic	0.95	0.78	0.77	0.097	7.94
Overall	0.89	0.66	0.68	0.076	8.95

Formulas (18.10)–(18.14) apply even when k, the number of categories, is equal to two. They may therefore be applied to the study of each category's reliability, as shown in Table 18.6 for the data of Table 18.1.

Note that the overall value of kappa is equal to the sum of the individual differences $p_o - p_e$ (i.e., of the numerators of the individual kappas) divided by the sum of the individual differences $1 - p_e$ (i.e., of the denominators of the individual kappas),

$$\hat{\bar{\kappa}} = \frac{(0.90 - 0.68) + (0.93 - 0.86) + (0.95 - 0.78)}{(1 - 0.68) + (1 - 0.86) + (1 - 0.78)} = \frac{0.46}{0.68} = 0.68,$$

confirming that $\hat{\bar{\kappa}}$ is a weighted average of the individual $\hat{\kappa}$'s.

For testing the hypothesis that the underlying value of kappa (either overall or for a single category) is equal to a prespecified value κ *other than zero*, Fleiss, Cohen, and Everitt (1969) showed that the appropriate standard error of $\hat{\kappa}$ is estimated by

$$\widehat{\text{se}}(\hat{\kappa}) = \frac{\sqrt{A + B - C}}{(1 - p_e)\sqrt{n}}, \tag{18.15}$$

where

$$A = \sum_{i=1}^{k} p_{ii}\left[1 - (p_{i.} + p_{.i})(1 - \hat{\kappa})\right]^2, \tag{18.16}$$

$$B = (1 - \hat{\kappa})^2 \sum\sum_{i \neq j} p_{ij}(p_{.i} + p_{j.})^2, \tag{18.17}$$

$$C = \left[\hat{\kappa} - p_e(1 - \hat{\kappa})\right]^2. \tag{18.18}$$

The hypothesis that κ is the underlying value would be rejected if the critical ratio

$$z = \frac{|\hat{\kappa} - \kappa|}{\widehat{\text{se}}(\hat{\kappa})} \tag{18.19}$$

were found to be significantly large from tables of the normal distribution.

An approximate $100(1-\alpha)\%$ confidence interval for κ is

$$\hat{\kappa} - z_{\alpha/2}\widehat{\text{se}}(\hat{\kappa}) \le \kappa \le \hat{\kappa} + z_{\alpha/2}\widehat{\text{se}}(\hat{\kappa}). \tag{18.20}$$

Consider testing the hypothesis that the overall value of kappa underlying the data in Table 18.1 is 0.80. The three quantities (18.16)–(18.18) needed to determine the standard error of $\hat{\bar{\kappa}}$ are

$$\begin{aligned}
A &= 0.75\left[1-(0.80+0.80)(1-0.68)\right]^2 \\
&\quad + 0.04\left[1-(0.10+0.05)(1-0.68)\right]^2 \\
&\quad + 0.10\left[1-(0.10+0.15)(1-0.68)\right]^2 \\
&= 0.2995, \\
B &= (1-0.68)^2\Big[0.01(0.80+0.10)^2 + 0.04(0.80+0.10)^2 \\
&\qquad + 0.05(0.05+0.80)^2 + 0.01(0.05+0.10)^2 \\
&\qquad + 0(0.15+0.80)^2 + 0(0.15+0.10)^2\Big] \\
&= 0.0079, \\
C &= \left[0.68-0.66(1-0.68)\right]^2 = 0.2198.
\end{aligned}$$

Thus

$$\widehat{\text{se}}(\hat{\bar{\kappa}}) = \frac{\sqrt{0.2995+0.0079-0.2198}}{(1-0.66)\sqrt{100}} = 0.087$$

and

$$z = \frac{|0.68-0.80|}{0.087} = 1.38,$$

so the hypothesis that $\bar{\kappa} = 0.80$ is not rejected.

Suppose one wishes to compare and combine g (≥ 2) independent estimates of kappa. The theory of Section 10.1 applies. Define, for the mth estimate, $V_m(\hat{\kappa}_m)$ to be the squared standard error of $\hat{\kappa}_m$, that is, the square of the expression in (18.15). The combined estimate of the supposed common value of kappa is, say,

$$\hat{\kappa}_{\text{overall}} = \frac{\displaystyle\sum_{m=1}^{g} \frac{\hat{\kappa}_m}{V_m(\hat{\kappa}_m)}}{\displaystyle\sum_{m=1}^{g} \frac{1}{V_m(\hat{\kappa}_m)}}. \tag{18.21}$$

To test the hypothesis that the g underlying values of kappa are equal, the value of

$$\chi^2_{\text{equal }\kappa\text{'s}} = \sum_{m=1}^{g} \frac{\left(\hat{\kappa}_m - \hat{\kappa}_{\text{overall}}\right)^2}{V_m\left(\hat{\kappa}_m\right)} \tag{18.22}$$

may be referred to tables of chi squared with $g-1$ df. The hypothesis is rejected if the value is significantly large. The limits of an approximate $100(1-\alpha)\%$ confidence interval for the supposed common underlying value are given by

$$\hat{\kappa}_{\text{overall}} \pm z_{\alpha/2} \sqrt{\frac{1}{\sum_{m=1}^{g} \frac{1}{V_m\left(\hat{\kappa}_m\right)}}}\,. \tag{18.23}$$

18.2. WEIGHTED KAPPA

Cohen (1968) (see also Spitzer et al. 1967) generalized his kappa measure of interrater agreement to the case where the relative seriousness of each possible disagreement could be quantified. Suppose that, independently of the data actually collected, agreement weights, say w_{ij} $(i = 1, \ldots, k;\ j = 1, \ldots, k)$, are assigned on rational or clinical grounds to the k^2 cells (see Cicchetti, 1976). The weights are restricted to lie in the interval $0 \le w_{ij} \le 1$ and to be such that

$$w_{ii} = 1 \tag{18.24}$$

(i.e., exact agreement is given maximal weight),

$$0 \le w_{ij} < 1 \qquad \text{for} \quad i \ne j \tag{18.25}$$

(i.e., all disagreements are given less than maximal weight), and

$$w_{ij} = w_{ji} \tag{18.26}$$

(i.e., the two raters are considered symmetrically).

The observed weighted proportion of agreement is, say,

$$p_{o(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij}, \tag{18.27}$$

where the proportions p_{ij} are arrayed as in Table 18.5, and the chance-expected weighted proportion of agreement is, say,

$$p_{e(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i.} p_{.j}. \tag{18.28}$$

Weighted kappa is then given by

$$\hat{\kappa}_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}. \tag{18.29}$$

Note that, when $w_{ij} = 0$ for all $i \neq j$ (i.e., when all disagreements are considered as being equally serious), then weighted kappa becomes identical to the overall kappa given in (18.12).

The interpretation of the magnitude of weighted kappa is like that of unweighted kappa: $\hat{\kappa}_w \geq 0.75$ or so signifies excellent agreement, for most purposes, and $\hat{\kappa}_w \leq 0.40$ or so signifies poor agreement.

Suppose that the k categories are ordered and that the decision is made to apply a two-way analysis of variance to the data resulting from taking the numerals $1, 2, \ldots, k$ as bona fide measurements. Bartko (1966) gives a formula for the intraclass correlation coefficient derived from this analysis of variance, and Fleiss and Cohen (1973) have shown that, aside from a term involving the factor $1/n$, the intraclass correlation coefficient is identical to weighted kappa provided the weights are taken as

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}. \tag{18.30}$$

Independently of Cohen (1968), Cicchetti and Allison (1971) proposed a statistic for measuring interrater reliability that is formally identical to weighted kappa. They suggested that the weights be taken as

$$w_{ij} = 1 - \frac{|i-j|}{k-1}. \tag{18.31}$$

The sampling distribution of weighted kappa was derived by Fleiss, Cohen, and Everitt (1969) and confirmed by Cicchetti and Fleiss (1977), Landis and Koch (1977a), Fleiss and Cicchetti (1978), and Hubert (1978). For testing the hypothesis that the underlying value of weighted kappa is zero, the appropriate estimated standard error of $\hat{\kappa}_w$ is

$$\widehat{\text{se}}_0(\hat{\kappa}_w) = \frac{1}{(1 - p_{e(w)})\sqrt{n}} \sqrt{\sum_{i=1}^{k} \sum_{j=1}^{k} p_{i.} p_{.j} \left[w_{ij} - (\overline{w}_{i.} + \overline{w}_{.j}) \right]^2 - p_{e(w)}^2}, \tag{18.32}$$

where

$$\overline{w}_{i.} = \sum_{j=1}^{k} p_{.j} w_{ij} \tag{18.33}$$

and

$$\overline{w}_{.j} = \sum_{i=1}^{k} p_{i.} w_{ij}, \tag{18.34}$$

The hypothesis may be tested by referring the value of the critical ratio

$$z = \frac{\hat{\kappa}_w}{\widehat{\text{se}}_0(\hat{\kappa}_w)} \tag{18.35}$$

to tables of the standard normal distribution.

For testing the hypothesis that the underlying value of weighted kappa is equal to a prespecified κ_w *other than zero*, the appropriate formula for the estimated standard error of $\hat{\kappa}_w$ is

$$\widehat{\text{se}}(\hat{\kappa}_w) = \frac{1}{(1 - p_{e(w)})\sqrt{n}} \times \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left[w_{ij} - (\bar{w}_{i.} + \bar{w}_{.j})(1 - \hat{\kappa}_w)\right]^2 - \left[\hat{\kappa}_w - p_{e(w)}(1 - \hat{\kappa}_w)\right]^2}. \tag{18.36}$$

The hypothesis may be tested by referring the value of the critical ratio

$$z = \frac{|\hat{\kappa}_w - \kappa_w|}{\widehat{\text{se}}(\hat{\kappa}_w)} \tag{18.37}$$

to tables of the standard normal distribution and rejecting the hypothesis if the critical ratio is too large.

It may be shown (see Problem 18.4) that the standard errors of unweighted kappa given in (18.13) and (18.15) are special cases of the standard errors of weighted kappa given in (18.32) and (18.36) when $w_{ii} = 1$ for all i and $w_{ij} = 0$ for all $i \neq j$.

Some attempts have been made to generalize kappa to the case where each subject is rated by each of the same set of more than two raters (Light, 1971; Landis and Koch, 1977a). Kairam et al. (1993) use the multivariate multiple noncentral hypergeometric distribution to study kappa in the case of $m \geq 2$ fixed raters with a prespecified interview schedule of subjects. Their analysis allows some subjects not to be seen by some raters. We consider in the next section the problem of different raters for different subjects when (i) $k = 2$ with varying m_i, or (ii) $k > 2$ with $m_i = m$ for all i. Kraemer (1980) considered the case in which $k > 2$ with varying m_i.

18.3. MULTIPLE RATINGS PER SUBJECT WITH DIFFERENT RATERS

Suppose that a sample of n subjects has been studied, with m_i being the number of ratings on the ith subject. The raters responsible for rating one

subject are not assumed to be same as those responsible for rating another. Suppose, further, that $k = 2$, that is, that the ratings consist of classifications into one of two categories; the case $k > 2$ will be considered later in this section. Finally, let x_i denote the number of (arbitrarily defined) positive ratings on subject i, so that $m_i - x_i$ is the number of negative ratings on him.

Identities between intraclass correlation coefficients and kappa statistics will be exploited to derive a kappa statistic by starting with an analysis of variance applied to the data (forming a one-way layout) obtained by coding a positive rating as 1 and a negative rating as 0. This was precisely the approach taken by Landis and Koch (1977b), except that they took the number of degrees of freedom for the mean square between subjects to be $n - 1$ instead of, as below, n.

Define the overall proportion of positive ratings to be

$$\bar{p} = \frac{\sum_{i=1}^{n} x_i}{n\bar{m}}, \tag{18.38}$$

where

$$\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n}, \tag{18.39}$$

the mean number of ratings per subject. If the number of subjects is large (say, $n \geq 20$), the mean square between subjects (BMS) is approximately equal to

$$\text{BMS} = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - m_i \bar{p})^2}{m_i} \tag{18.40}$$

and the mean square within subjects (WMS) is equal to

$$\text{WMS} = \frac{1}{n(\bar{m} - 1)} \sum_{i=1}^{n} \frac{x_i(m_i - x_i)}{m_i}. \tag{18.41}$$

Technically, the intraclass correlation coefficient should be estimated as

$$r = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (m_0 - 1)\text{WMS}}, \tag{18.42}$$

where

$$m_0 = \bar{m} - \frac{\sum_{i=1}^{n} (m_i - \bar{m})^2}{n(n-1)\bar{m}}. \tag{18.43}$$

If n is at all large, though, m_0 and $\bar{m}$ will be very close in magnitude. If m_0 is replaced by $\bar{m}$ in (18.42), the resulting expression for the intraclass

correlation coefficient, and therefore for kappa, is

$$\hat{\kappa} = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (\overline{m} - 1)\text{WMS}}$$

$$= 1 - \frac{\sum_{i=1}^{n} \frac{x_i(m_i - x_i)}{m_i}}{n(\overline{m} - 1)\bar{p}\bar{q}}, \qquad (18.44)$$

where $\bar{q} = 1 - \bar{p}$.

$\hat{\kappa}$ has the following properties. If there is no subject-to-subject variation in the proportion of positive ratings (i.e., if $x_i/m_i = \bar{p}$ for all i, with $\bar{p}$ not equal to either 0 or 1), then there is more disagreement within subjects than between subjects. In this case $\hat{\kappa}$ may be seen to assume its minimum value of $-1/(\overline{m} - 1)$.

If the several proportions x_i/m_i vary exactly as binomial proportions with parameters m_i and a common probability $\bar{p}$, then there is as much similarity within subjects as between subjects. In this case, the value of $\hat{\kappa}$ is equal to 0.

If each proportion x_i/m_i assumes either the values 0 or 1, then there is perfect agreement within subjects. In this case, $\hat{\kappa}$ may be seen to assume the value 1.

Consider the hypothetical data of Table 18.7 on $n = 25$ subjects. For these data, the mean number of ratings per subject is

$$\overline{m} = \frac{81}{25} = 3.24,$$

Table 18.7. ***Hypothetical ratings by different sets of raters on n = 25 subjects***

Subject i	Number of Raters, m_i	Number of Positive Ratings, x_i	i	m_i	x_i
1	2	2	14	4	3
2	2	0	15	2	0
3	3	2	16	2	2
4	4	3	17	3	1
5	3	3	18	2	1
6	4	1	19	4	1
7	3	0	20	5	4
8	5	0	21	3	2
9	2	0	22	4	0
10	4	4	23	3	0
11	5	5	24	3	3
12	3	3	25	2	2
13	4	4	Total	81	46

the overall proportion of positive ratings is

$$\bar{p} = \frac{46}{25 \times 3.24} = 0.568,$$

and the value of $\sum x_i(m_i - x_i)/m_i$ is

$$\sum_{i=1}^{25} \frac{x_i(m_i - x_i)}{m_i} = 6.30.$$

The value of kappa in (18.44) for these ratings is therefore

$$\begin{aligned}\hat{\kappa} &= 1 - \frac{6.30}{25(3.24 - 1) \times 0.568 \times 0.432} \\ &= 0.54,\end{aligned}$$

indicating only a modest degree of interrater agreement.

Fleiss and Cuzick (1979) derived the standard error of $\hat{\kappa}$ appropriate for testing the hypothesis that the underlying value of kappa is 0. Define $\overline{m}_H$ to be the *harmonic mean* of the number of ratings per subject, that is,

$$\overline{m}_H = \frac{n}{\sum_{i=1}^{n} 1/m_i}. \tag{18.45}$$

The standard error of $\hat{\kappa}$ is estimated by

$$\widehat{\text{se}}_0(\hat{\kappa}) = \frac{1}{(\overline{m} - 1)\sqrt{n\overline{m}_H}} \sqrt{2(\overline{m}_H - 1) + \frac{(\overline{m} - \overline{m}_H)(1 - 4\bar{p}\bar{q})}{\overline{m}\bar{p}\bar{q}}}, \tag{18.46}$$

and the hypothesis may be tested by referring the value of the critical ratio

$$z = \frac{\hat{\kappa}}{\widehat{\text{se}}_0(\hat{\kappa})} \tag{18.47}$$

to tables of the standard normal distribution.

For the data of Table 18.7,

$$\overline{m}_H = \frac{25}{8.5167} = 2.935$$

and

$$\widehat{\text{se}}_0(\hat{\kappa}) = \frac{1}{(3.24-1)\sqrt{25\times 2.935}} \times \sqrt{2(2.935-1) + \frac{(3.24-2.935)(1-4\times 0.568\times 0.432)}{3.24\times 0.568\times 0.432}} = 0.103.$$

The value of the critical ratio in (18.47) is then

$$z = \frac{0.54}{0.103} = 5.24,$$

indicating that $\hat{\kappa}$ is significantly greater than zero.

Suppose, now, that the number of categories into which ratings are made is $k \geq 2$. Denote by $\bar{p}_j$ the overall proportion of ratings in category j and by $\hat{\kappa}_j$ the value of kappa for category, $j, j = 1, \ldots, k$. Landis and Koch (1977b) proposed taking the weighted average

$$\hat{\bar{\kappa}} = \frac{\sum_{j=1}^{k} \bar{p}_j \bar{q}_j \hat{\kappa}_j}{\sum_{j=1}^{k} \bar{p}_j \bar{q}_j} \tag{18.48}$$

as an overall measure of interrater agreement, where $\bar{q}_j = 1 - \bar{p}_j$. The standard error of $\hat{\bar{\kappa}}$ has yet to be derived, when the numbers of ratings per subject vary, to test the hypothesis that the underlying value is zero.

When, however, the number of ratings per subject is constant and equal to m, simple expressions for $\hat{\kappa}_j$, $\hat{\bar{\kappa}}$, and their standard errors are available. Define x_{ij} to be the number of ratings on subject i $(i = 1, \ldots, n)$ into category j $(j = 1, \ldots, k)$; note that

$$\sum_{j=1}^{k} x_{ij} = m \tag{18.49}$$

for all i. The value of $\hat{\kappa}_j$ is then

$$\hat{\kappa}_j = 1 - \frac{\sum_{i=1}^{n} x_{ij}(m - x_{ij})}{nm(m-1)\bar{p}_j \bar{q}_j}, \tag{18.50}$$

and the value of $\hat{\bar{\kappa}}$ is

$$\hat{\bar{\kappa}} = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2}{nm(m-1)\sum_{j=1}^{k} \bar{p}_j \bar{q}_j}. \tag{18.51}$$

Table 18.8. ***Five ratings on each of ten subjects into one of three categories***

Subject	Number of Ratings into Category			$\sum_{j=1}^{3} x_{ij}^2$
	1	2	3	
1	1	4	0	17
2	2	0	3	13
3	0	0	5	25
4	4	0	1	17
5	3	0	2	13
6	1	4	0	17
7	5	0	0	25
8	0	4	1	17
9	1	0	4	17
10	3	0	2	13
Total	20	12	18	174

Algebraically equivalent versions of these formulas were first presented by Fleiss (1971), who showed explicitly how they represent chance-corrected measures of agreement.

Table 18.8 presents hypothetical data representing, for each of $n = 10$ subjects, $m = 5$ ratings into one of $k = 3$ categories.

The three overall proportions are $\bar{p}_1 = 20/50 = 0.40$, $\bar{p}_2 = 12/50 = 0.24$, and $\bar{p}_3 = 18/50 = 0.36$. For category 1, the numerator in expression (18.50) for $\hat{\kappa}_1$ is

$$\sum_{i=1}^{10} x_{i1}(5 - x_{i1}) = 1 \times (5-1) + 2 \times (5-2) + \cdots + 3 \times (5-3) = 34,$$

and thus

$$\hat{\kappa}_1 = 1 - \frac{34}{10 \times 5 \times 4 \times 0.40 \times 0.60} = 0.29.$$

Similarly, $\hat{\kappa}_2 = 0.67$ and $\hat{\kappa}_3 = 0.35$. The overall value of $\hat{\bar{\kappa}}$ is, by (18.51),

$$\hat{\bar{\kappa}} = 1 - \frac{10 \times 25 - 174}{10 \times 5 \times 4 \times (0.40 \times 0.60 + 0.24 \times 0.76 + 0.36 \times 0.64)} = 0.42.$$

Alternatively,

$$\hat{\bar{\kappa}} = \frac{(0.40 \times 0.60) \times 0.29 + (0.24 \times 0.76) \times 0.67 + (0.36 \times 0.64) \times 0.35}{0.40 \times 0.60 + 0.24 \times 0.76 + 0.36 \times 0.64}$$

$$= 0.42.$$

When the numbers of ratings per subject are equal, Fleiss, Nee, and Landis (1979) derived and confirmed the following formulas for the approximate standard errors of $\hat{\bar{\kappa}}$ and $\hat{\kappa}_j$, each appropriate for testing the hypothesis that the underlying value is zero:

$$\widehat{\mathrm{se}}_0(\hat{\bar{\kappa}}) = \frac{\sqrt{2}}{\sum_{j=1}^{k} \bar{p}_j \bar{q}_j \sqrt{nm(m-1)}} \times \sqrt{\left(\sum_{j=1}^{k} \bar{p}_j \bar{q}_j \right)^2 - \sum_{j=1}^{k} \bar{p}_j \bar{q}_j \left(\bar{q}_j - \bar{p}_j \right)}, \tag{18.52}$$

and

$$\mathrm{se}_0(\hat{\kappa}_j) = \sqrt{\frac{2}{nm(m-1)}}. \tag{18.53}$$

Note that $\mathrm{se}_0(\hat{\kappa}_j)$ is independent of $\bar{p}_j$ and $\bar{q}_j$! Further, it is easily checked that formula (18.53) is a special case of (18.46) when the m_i's are all equal, because then $\overline{m} = \overline{m}_H = m$.

For the data of Table 18.8,

$$\sum_{j=1}^{3} \bar{p}_j \bar{q}_j = 0.40 \times 0.60 + 0.24 \times 0.76 + 0.36 \times 0.64 = 0.6528$$

and

$$\begin{aligned} \sum_{j=1}^{3} \bar{p}_j \bar{q}_j \left(\bar{q}_j - \bar{p}_j \right) &= 0.40 \times 0.60 \times (0.60 - 0.40) + 0.24 \times 0.76 \times (0.76 - 0.24) \\ &\quad + 0.36 \times 0.64 \times (0.64 - 0.36) \\ &= 0.2074, \end{aligned}$$

so that

$$\widehat{\mathrm{se}}_0(\hat{\bar{\kappa}}) = \frac{\sqrt{2}}{0.6528\sqrt{10 \times 5 \times 4}} \sqrt{0.6528^2 - 0.2074} = 0.072.$$

Because

$$z = \frac{\hat{\bar{\kappa}}}{\widehat{\mathrm{se}}_0(\hat{\bar{\kappa}})} = \frac{0.42}{0.072} = 5.83,$$

the overall value of kappa is significantly different from zero (although its magnitude indicates only mediocre reliability).

The approximate standard error of each $\hat{\kappa}_j$ is, by (18.53),

$$\mathrm{se}_0(\hat{\kappa}_j) = \sqrt{\frac{2}{10 \times 5 \times 4}} = 0.10.$$

Each individual kappa is significantly different ($p < 0.01$) from zero, but only $\hat{\kappa}_2$ approaches a value suggestive of fair reliability.

Various approaches have been taken to obtain the standard error of κ. Fleiss and Davies (1982) and Bloch and Kraemer (1989) obtain an asymptotic variance, and a jackknife technique is proposed by Fleiss and Davies (1982), Schouten (1986), and Flack (1987). Flack (1987) proposes a skewness-corrected confidence interval using a jackknife estimate of the third moment of the distribution of delete-one κ statistics. Donner and Eliasziw (1992) obtain a standard error with a method based on a goodness-of-fit test statistic frequently used for clustered binary data. Lee and Tu (1994) propose yet another confidence interval for κ in the case of two raters with binary ratings, by reparameterizing κ as a monotone function of p_{11}. Garner (1991) obtains the standard error conditioning on the margins. Hale and Fleiss (1993) give two variance estimates of κ depending on whether the rater effect is treated as fixed or random. Lipsitz, Laird, and Brennan (1994) provide an asymptotic variance of κ statistics based on the theory of estimating equations.

18.4. FURTHER APPLICATIONS

Even though the various kappa statistics were originally developed and were illustrated here for the measurement of interrater agreement, their applicability extends far beyond this specific problem. In fact, they are useful for measuring, on categorical data, such constructs as "similarity," "concordance," and "clustering." Some examples will be given.

1. In a study of the correlates or determinants of drug use among teenagers, it may be of interest to determine how concordant the attitudes toward drug use are between each subject's same-sex parent and the subject's best friend. Either unweighted kappa or weighted kappa (Section 18.1) may be used, with rater A replaced by parent and rater B by best friend.

2. Suppose that m monitoring stations are set up in a city to measure levels of various pollutants and that, on each of n days, each station is characterized by whether or not the level of a specified pollutant (e.g., sulfur dioxide) exceeds an officially designated threshold. The version of kappa presented in Section 18.3 may be applied to describe how well (or poorly) the several stations agree.

3. Consider a study of the role of familial factors in the development of a condition such as adolescent hypertension. Suppose that n sibships are

studied and that m_i is the number of siblings in the ith sibship. The version of kappa presented in Section 18.3 may be applied to describe the degree to which there is familial aggregation in the condition.

4. Many of the indices of agreement cited in Section 18.1 are used in numerical taxonomy (Sneath and Sokal, 1973) to describe the degree of similarity between different study units; in fact, p_s (18.2) was originally proposed for this purpose by Dice (1945). Suppose that two units (people, languages, or whatever) are being compared with respect to whether they possess or do not possess each of n dichotomous characteristics. The proportions a–d in the left-hand part of Table 18.2 then refer to the proportion of all n characteristics that both units possess, the proportion that one possesses but the other does not, and so on. Corrections for chance-expected similarity in this kind of problem are as important as corrections for chance-expected agreement in the case of interrater reliability. Bloch and Kraemer (1989) discuss kappa as a measure of agreement and association.

5. Studies in which several controls are matched with each case or each experimental unit were discussed in Section 13.3. If the several controls in each matched set were successfully matched, the responses by the controls from the same set should be more similar than the responses by controls from different sets. The version of kappa presented in Section 18.2 may be used to describe how successful the matching was.

6. Although κ is widely used in psychology and educational research, its application extends to periodontal research (Boushka et al., 1990), econometrics (Hirschberg and Slottje, 1989), veterinary epidemiology (Shourkri, Martin, and Mian, 1995), anesthesiology (Posner et al., 1990), neurology (Kairam et al., 1993), and radiology (Musch et al., 1984).

Whether used to measure agreement, or, more generally, similarity, kappa in effect treats all the raters or units symmetrically. When one or more of the sources of ratings may be viewed as a standard, however (two of $m = 5$ raters, e.g., may be senior to the others, or one of the air pollution monitoring stations in example 2 may employ more precise measuring instruments than the others), kappa may no longer be appropriate, and the procedures described by Light (1971), Williams (1976), and Wackerley, McClave, and Rao (1978) should be employed instead.

18.5.* INTERRATER AGREEMENT AS ASSOCIATION IN A MULTIVARIATE BINARY VECTOR

Many problems of interrater agreement can be solved in the framework of clustered categorical data (see Chapter 15). For a binary rating, the notion of interrater agreement is closely related to the correlation among the binary ratings clustered within a subject. Specifically, suppose there are m_i raters, each of whom gives a two-category rating to subject i for $i = 1, \ldots, n$. Let the

binary indicator Y_{ij} be 1 if rater j judges subject i positive, and 0 if negative, for $j = 1, \ldots, m_i$. Then $Y_i = (Y_{i1}, \ldots, Y_{im_i})'$ constitutes a vector of binary outcomes, and the dependence among its components can be characterized by the intraclass correlation coefficient (ICC) or kappa, among many other measures. When m_i is the same for all i, the ICC and κ are identical.

One way to specify the distribution of the Y_{ij}'s is to consider all possible 2^{m_i} mutually exclusive response profiles and assume a 2^{m_i}-variate multinomial distribution. Some authors specify the multivariate distribution of Y_i this way, while some focus on the distribution of the total number of positive ratings for subject i, Y_{i+}, and assume it has a beta-binomial distribution; in either case they express kappa in terms of the parameters of the chosen distribution and obtain the maximum likelihood estimate (mle). See Verducci, Mack, and DeGroot (1988), Shoukri, Martin, and Mian (1995), Shoukri and Mian (1995), and Barlow (1996). Other authors construct a multivariate distribution using a latent class model; see Aickin (1990), Agresti and Lang (1993), and Uebersax (1993).

In a different approach, the pairwise association between Y_{ij} and Y_{ik} can be expressed as a function of kappa without making a full distributional assumption. Landis and Koch (1977b) structure the correlation using a random effects model. They assume

$$Y_{ij} = P + s_i + e_{ij},$$

where P is the probability of a positive rating, the s_i's are independent and identically distributed with mean 0 and variance σ_s^2, the e_{ij}'s are similarly distributed with mean 0 and variance σ_e^2, and the s_i's and e_{ij}'s are mutually independent. Then Y_{ij} and Y_{ik} are conditionally independent given the random effect s_i which is unique to subject i, but are marginally correlated, because they share the random effect s_i. See Section 15.5.2 at expression (15.42). The intraclass correlation coefficient is

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

The authors use a moment estimator to estimate ρ and derive its standard error.

Lipsitz, Laird, and Brennan (1994) propose a class of estimators for kappa using an estimating-equation approach (see Section 15.5.1). Assuming that each subject has the same probability of a positive rating, say $P = P(Y_{ij} = 1)$, and the same joint probability of being rated positive by a pair of raters for all pairs of raters, $P_{11} = E(Y_{ij}Y_{ik}) = P(Y_{ij} = 1, Y_{ik} = 1)$, kappa can be written as a function of the probability of agreement under two assumptions: nonindependence among the elements of Y_i, and independence. The probability of agreement without assuming independence, P_a, is

$$\begin{aligned} P_a &= P(Y_{ij} = 1, Y_{ik} = 1) + P(Y_{ij} = 0, Y_{ik} = 0) \\ &= P_{11} + \{(1 - P) - (P - P_{11})\} = P_{11} + 1 - 2P + P_{11}. \end{aligned} \quad (18.54)$$

The chance-expected probability of agreement, P_e, is the probability of agreement under marginal independence among the elements in Y_i:

$$P_e = P^2 + (1-P)^2. \tag{18.55}$$

With

$$\kappa = \frac{P_a - P_e}{1 - P_e} = 1 - \frac{1 - P_a}{1 - P_e},$$

after substitution of (18.54) and (18.55) we have

$$\kappa = 1 - \frac{P - P_{11}}{P(1-P)}. \tag{18.56}$$

We can rewrite P_{11} in terms of P and κ thus:

$$P_{11} = P^2 + \kappa P(1-P).$$

Lipsitz, Laird, and Brennan (1994) construct a class of estimating equations each of whose solutions becomes an estimate of kappa. Based on the identities $E(Y_{i+}) = m_i P$ and $E\{Y_{i+}(Y_{i+}-1)\} = P_{11} m_i (m_i - 1)$, the authors construct a joint estimating equation,

$$\begin{pmatrix} U_1(P) \\ U_2(\kappa, P) \end{pmatrix} = 0$$

with

$$U_1(P) = \sum_{i=1}^{n} \frac{Y_{i+} - m_i P}{v_i},$$

and

$$U_2(\kappa, P) = \sum_{i=1}^{n} \frac{Y_{i+}(Y_{i+}-1) - P_{11} m_i (m_i - 1)}{w_i},$$

where v_i and w_i are weights to be chosen. The estimating equation is unbiased, that is, $E\{U_1(P)\} = E\{U_2(\kappa, P)\} = 0$ for all κ and P, and, as explained in Section 15.5.1, the solution is consistent and asymptotically normal. Applying further results from the standard theory of estimating equations, the variance of $\hat{\kappa}$ has a sandwich-type estimator which can be obtained easily. A convenience of this approach is that on choosing the weights v_i and w_i appropriately, the solution of the estimating equation coincides with existing kappa statistics, including the kappa statistic of Fleiss

(1971) and the weighted kappa statistic of Schouten (1986). For example, Fleiss' kappa can be obtained by solving

$$U_1(\hat{P}) = \sum_{i=1}^{n} U_{i1} = \sum_{i=1}^{n} \left(Y_{i+} - m_i \hat{P}\right) = 0 \tag{18.57}$$

and

$$-U_2(\hat{\kappa}, \hat{P}) = \sum_{i=1}^{n} U_{2i} = \sum_{i=1}^{n} \left\{ \frac{Y_{i+}(m_i - Y_{i+})}{m_i - 1} - (1 - \hat{\kappa})\hat{P}(1 - \hat{P})m_i \right\}$$
$$= 0. \tag{18.58}$$

The sandwich-type variance of Lipsitz, Laird, and Brennan (1994) is asymptotically equivalent to the jackknife variance estimate proposed by Schouten (1986). The sandwich variance of Fleiss' kappa statistic has the form $\mathrm{Var}(\hat{\kappa}) = \sum_i V_i^2$, where

$$V_i = \frac{U_{2i} - \dfrac{(1-\kappa)(1-2P)}{\overline{m}} U_{1i}}{nP(1-P)}.$$

The authors also show that the asymptotic relative efficiency against the mle assuming a beta-binomial distribution (Verducci, Mack, and DeGroot, 1988) is highest for Fleiss' kappa, lower for weighted kappa (Schouten, 1986), and lowest for unweighted kappa, where both v_i and w_i are constants.

The estimating-equation approach can be extended to the regression case in which kappa is modeled as a function of covariates. Alternative ways of incorporating covariates and testing homogeneity of kappa across covariate levels are discussed by Barlow, Lai, and Azen (1991), Barlow (1996), and Donner, Eliasziew, and Klar (1996).

Both mle and estimating-equation estimators require a large sample size for inferences to be valid. Small-sample properties of kappa estimates have been studied by Koval and Blackman (1996) and Gross (1986). Lau (1993) provides higher-order kappa-type statistics for a dichotomous attribute with multiple raters.

Several authors investigate alternative measures of agreement. Kupper and Hafner (1989) discuss correcting for chance agreement when the raters' attribute selection probabilities are equal, and use a hypergeometric distribution. O'Connell and Dobson (1984) describe a class of agreement measures in which kappa is a special case. Uebersax (1993) considers a measure of agreement based on a latent-class model. Aickin (1990) uses a mixture of distributions assuming independent ratings and perfect agreement, and takes the mixing probability as a measure of agreement. He finds that his measure of agreement has a kappa-like form, but tends to be larger than Cohen's

kappa except in the case of uniform margins. Agresti (1992) and Banerjee, Capozzoli, and McSweeney (1999) give a review of measures of agreement, and Smeeton (1985) describes the early history of kappa.

PROBLEMS

18.1. Prove that, when each of the indices of agreement given by (18.1), (18.2), (18.4), (18.6), and (18.7) is corrected for chance-expected agreement using formula (18.8), the same formula for kappa (18.9) is obtained.

18.2. Prove that, when $k = 2$, the square of the critical ratio given in (18.14) is identical to the standard chi squared statistic without the continuity correction.

18.3. Suppose that $g = 3$ independent reliability studies of a given kind of rating have been conducted, with results as follows:

Study 1 ($n = 20$)			Study 2 ($n = 20$)			Study 3 ($n = 30$)		
	Rater B			Rater D			Rater F	
Rater A	+	−	Rater C	+	−	Rater E	+	−
+	0.60	0.05	+	0.75	0.10	+	0.50	0.20
−	0.20	0.15	−	0.05	0.10	−	0.10	0.20

(a) What are the three values of kappa? What are their standard errors [see (18.15)]? What is the overall value of kappa [see (18.21)]?

(b) Are the three estimates of kappa significantly different? [Refer the value of the statistic in (18.22) to tables of chi squared with 2 df.]

(c) Using (18.23), find an approximate 95% confidence interval for the common value of kappa.

18.4. Prove that, when $w_{ii} = 1$ for all i and $w_{ij} = 0$ for all $i \neq j$, the standard-error formulas (18.13) and (18.32) are identical. Prove that, with this same system of agreement weights, the standard-error formulas (18.15) and (18.36) are identical.

18.5. Prove that, when $k = 2$, formulas (18.52) and (18.53) are identical.

REFERENCES

Agresti, A. (1992). Modeling patterns of agreement and disagreement. *Statist. Methods Med. Res.*, **1**, 201–218.

Agresti, A. and Lang, J. B. (1993). Quasi-symmetrical latent class models, with application to rater agreement. *Biometrics*, **49**, 131–139.

Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics*, **46**, 293–302.

Armitage, P., Blendis, L. M., and Smyllie, H. C. (1966). The measurement of observer disagreement in the recording of signs. *J. R. Statist. Soc., Ser. A*, **129**, 98–109.

Armstrong, B. (1985). Measurement error in the generalised linear model. *Comm. Statist.*, **B14**, 529–544.

Banerjee, M., Capozzoli, M., and McSweeney, L. (1999). Beyond kappa: A review of interrater agreement measures. *Canad. J. Statist.*, **27**, 3–23.

Barlow, W. (1996). Measurement of interrater agreement with adjustment for covariates. *Biometrics*, **52**, 695–702.

Barlow, W., Lai, M.-Y., and Azen, S. P. (1991). A comparison of methods for calculating a stratified kappa. *Statist. in Med.*, **10**, 1465–1472.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychol. Rep.*, **19**, 3–11.

Bennett, B. M. (1972). Measures for clinicians' disagreements over signs. *Biometrics*, **28**, 607–612.

Bloch, D. A. and Kraemer, H. C. (1989). 2×2 kappa coefficients: Measures of agreement or association (C/R: pp. 1329–1330). *Biometrics*, **45**, 269–287.

Boushka, W. M., Martinez, Y. N., Prihoda, T. J., Dunford, R., and Barnwell, G. M. (1990). A computer program for calculating kappa: Application to interexaminer agreement in periodontal research. *Computer Methods Programs Biomed.*, **33**, 35–41.

Cicchetti, D. V. (1976). Assessing inter-rater reliability for rating scales: Resolving some basic issues. *Brit. J. Psychiatry*, **129**, 452–456.

Cicchetti, D. V. and Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *Am. J. EEG Technol.*, **11**, 101–109.

Cicchetti, D. V. and Fleiss, J. L. (1977). Comparison of the null distributions of weighted kappa and the C ordinal statistic. *Appl. Psychol. Meas.*, **1**, 195–201.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, **20**, 37–46.

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

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, **26**, 297–302.

Donner, A. and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval construction, significance-testing, and sample size estimation. *Statist. in Med.*, **11**, 1511–1519.

Donner, A., Eliasziew, M., and Klar, N. (1996). Testing the homogeneity of kappa statistics. *Biometrics*, **52**, 176–183.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika*, **16**, 407–424.

Flack, V. F. (1987). Confidence intervals for the interrater agreement measure kappa. *Comm. Statist.*, **A16**, 953–968.

Fleiss, J. L. (1965). Estimating the accuracy of dichotomous judgments. *Psychometrika*, **30**, 469–479.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.*, **76**, 378–382.

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

Fleiss, J. L. and Cicchetti, D. V. (1978). Inference about weighted kappa in the non-null case. *Appl. Psychol. Meas.*, **2**, 113–117.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.*, **33**, 613–619.

Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychol. Bull.*, **72**, 323–327.

Fleiss, J. L. and Cuzick, J. (1979). The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Appl. Psychol. Meas.*, **3**, 537–542.

Fleiss, J. L. and Davies, M. (1982). Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *Am. J. Epidemiol.*, **115**, 841–845.

Fleiss, J. L., Nee, J. C. M., and Landis, J. R. (1979). The large sample variance of kappa in the case of different sets of raters. *Psychol. Bull.*, **86**, 974–977.

Garner, J. B. (1991). The standard error of Cohen's kappa. *Statist. in Med.*, **10**, 767–775.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *J. Am. Statist. Assoc.*, **49**, 732–764.

Gross, S. T. (1986). The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics*, **42**, 883–893.

Hale, C. A. and Fleiss, J. L. (1993). Interval estimation under two study designs for kappa with binary classifications. *Biometrics*, **49**, 523–533.

Hirschberg, J. G. and Slottje, D. J. (1989). Remembrance of things past the distribution of earnings across occupations and the kappa criterion. *J. Econometrics*, **42**, 121–130.

Holley, J. W. and Guilford, J. P. (1964). A note on the *G* index of agreement. *Educ. Psychol. Meas.*, **32**, 281–288.

Hubert, L. J. (1978). A general formula for the variance of Cohen's weighted kappa. *Psychol. Bull.*, **85**, 183–184.

Kairam, R., Kline, J., Levin, B., Brambilla, D., Coulter, D., Kuban, K., Lansky, L., Marshall, P., Velez-Borras, J., and Rodriguez, E. (1993). Reliability of neurologic assessment in a collaborative study of HIV infection in children. *Ann. N.Y. Acad. Sci.*, **693**, 123–140.

Koval, J. J. and Blackman, N. J.-M. (1996). Estimators of kappa—exact small sample properties. *J. Statist. Comput. Simul.* **55**, 315–336.

Kraemer, H. C. (1980). Extension of the kappa coefficient. *Biometrics*, **36**, 207–216.

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. Pp. 139–150 in E. F. Borgatta (Ed.). *Sociological methodology 1970*. San Francisco: Jossey-Bass.

Kupper, L. L. and Hafner, K. B. (1989). On assessing interrater agreement for multiple attribute responses. *Biometrics*, **45**, 957–967.

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

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

Lau, T.-S. (1993). Higher-order kappa-type statistics for a dichotomous attribute in multiple ratings. *Biometrics*, **49**, 535–542.

Lee, J. J. and Tu, Z. N. (1994). A better confidence interval for kappa (κ) on measuring agreement between two raters with binary outcomes. *J. Comput. Graph. Statist.*, **3**, 301–321.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychol. Bull.*, **76**, 365–377.

Lipsitz, S. R., Laird, N. M., and Brennan, T. A. (1994). Simple moment estimates of the κ-coefficient and its variance. *Appl. Statist.*, **43**, 309–323.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *Brit. J. Psychiatry*, **130**, 79–83.

Maxwell, A. E. and Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. *Brit. J. Math. Statist. Psychol.*, **21**, 105–116.

Musch, D. C., Landis, J. R., Higgins, I. T. T., Gilson, J. C., and Jones, R. N. (1984). An application of kappa-type analysis to interobserver variation in classifying chest radiographs for pneumoconiosis. *Statist. in Med.*, **3**, 73–83.

O'Connell, D. L. and Dobson, A. J. (1984). General observer-agreement measures on individual subjects and groups of subjects. *Biometrics*, **40**, 973–983.

Posner, K. L., Sampson, P. D., Caplan, R. A., Ward, R. J., and Cheney, F. W. (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Statist. in Med.*, **9**, 1103–1115.

Rogot, E. and Goldberg, I. D. (1966). A proposed index for measuring agreement in test-retest studies. *J. Chronic Dis.*, **19**, 991–1006.

Schouten, H. J. A. (1986). Nominal scale agreement among observers. *Psychometrika*, **51**, 453–466.

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

Shoukri, M. M., Martin, S. W., and Mian, I. U. H. (1995). Maximum likelihood estimation of the kappa coefficient from models of matched binary responses. *Statist. in Med.*, **14**, 83–99.

Shoukri, M. M. and Mian, I. U. H. (1995). Maximum likelihood estimation of the kappa coefficient from logistic regression. *Statist. in Med.*, **15**, 1409–1419.

Smeeton, N. C. (1985). Early history of the kappa statistic. *Biometrics*, **41**, 795.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy*. San Francisco: W. H. Freeman.

Spitzer, R. L., Cohen, J., Fleiss, J. L., and Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis. *Arch. Gen. Psychiatry*, **17**, 83–87.

Spitzer, R. L. and Fleiss, J. L. (1974). A reanalysis of the reliability of psychiatric diagnosis. *Brit. J. Psychiatry*, **125**, 341–347.

Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical-treatment appropriateness. *J. Am. Statist. Assoc.*, **88**, 421–427.

Verducci, J. S., Mack, M. E., and DeGroot, M. H. (1988). Estimating multiple rater agreement for a rare diagnosis. *J. Multivariate Anal.*, **27**, 512–535.

Wackerley, D. D., McClave, J. T., and Rao, P. V. (1978). Measuring nominal scale agreement between a judge and a known standard. *Psychometrika*, **43**, 213–223.

Williams, G. W. (1976). Comparing the joint agreement of several raters with another rater. *Biometrics*, **32**, 619–627.