A review of statistical methods in the analysis of data arising from observer reliability studies (Part II)*

by J. RICHARD LANDIS** and GARY G. KOCH***

4 Methods proposed for nominal and ordinal data

Many research designs in studies of observer reliability give rise to categorical data via nomial scales (*e.g.*, states of mental health such as normal, neurosis, and depression) or ordinal scales (*e.g.*, stages of disease such as mild, moderate, and severe). In these situations, each of the d observers classifies each subject once into exactly one of a fixed set of L categories. As such, these designs are directly analogous to those giving rise to the standard ANOVA models in (2.1), (2.5), and (2.10) when the measurement scale is assumed to be quantitative. However, standard ANOVA procedures are rarely appropriate for the analysis of nominal and ordinal scaled data. As a result, these data are usually cross-classified into an L^d contingency table, and can then be analyzed by techniques developed for multidimensional contingency tables.

4.1 *Measures of association between two observers*

When each of $d = 2$ observers separately classifies n subjects on an L-point scale, the resulting data can be summarized in the $L \times L$ table of observed proportions shown in Table 6. In this case, $p_{kk'}$ is the proportion of subjects classified into category k by observer 1 and into category k' by observer 2. Moreover, the diagonal elements $\{p_{kk}\}$ for $k = 1, 2, ..., L$ represent the proportions of the subjects classified into each of the respective agreement category combinations.

Various indices which characterize the association between the row and column classifications have been proposed for $L \times L$ contingency tables. For example KENDALL and STUART [57] discussed a coefficient of contingency due to Pearson denoted by

$$P = \left\{\frac{\chi^2}{n+\chi^2}\right\}^{\frac{1}{2}} \tag{4.1}$$

where χ^2 is the Pearson chi-square statistic for independence. This P coefficient ranges from 0 (for complete independence) to an upper limit of $((L-1)/L)^{\frac{1}{2}}$ (for perfect agreement) between the two observers. As such, the upper limit of this coefficient depends on the number of categories in the measurement scale.

In order to avoid this undesirable scale-dependency property of P in (4.1),

---

* Part I appeared in Statistica Neerlandica, nr. 3, 1975.

** Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48104, U.S.A.

*** Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, North Carolina 27514, U.S.A.

Table 6. Observed proportions resulting from two observers classifying n subjects on an L-point scale

		observer 2				
		1	2	...	L	total
observer 1	1	p_{11}	p_{12}	...	p_{1L}	$p_{1\cdot}$
	2	p_{21}	p_{22}	...	p_{2L}	$p_{2\cdot}$
	L	p_{L1}	p_{L2}	...	p_{LL}	$p_{L\cdot}$
	total	$p_{\cdot 1}$	$p_{\cdot 2}$	...	$p_{\cdot L}$	1

TSCHUPROW proposed an alternative function of χ^2 for the $L \times L$ table, which is given in KENDALL and STUART [57] as

$$T = \left\{ \frac{\chi^2}{n(L-1)} \right\}^{\frac{1}{2}}. \tag{4.2}$$

T ranges from 0 (for complete independence) to +1 (for perfect agreement) between the two observers. In this regard, T is a natural extension of ϕ in (3.2), since $T = \phi$ when $L = 2$.

4.2 *Measures of agreement between two observers*

As discussed in Section 3.2, agreement is a special case of association which reflects the extent to which observers classify a given subject identically into the same category. For this purpose, the most elementary index of agreement is based on the proportion of the subjects classified into the same category by the two observers, and can be estimated by

$$p_o = \sum_{k=1}^{L} p_{kk}, \tag{4.3}$$

which is a direct extension of the index of crude agreement in (3.3). Under the baseline constraints of complete independence between ratings by the two observers, the expected agreement proportion corresponding to (4.3) is estimated by

$$p_e = \sum_{k=1}^{L} p_{k\cdot} p_{\cdot k}. \tag{4.4}$$

Moreover, COHEN [21] proposed a standardized coefficient of agreement for nominal scales in terms of (4.3) and (4.4) which can be estimated by

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e}. \tag{4.5}$$

As defined here, $\hat{\kappa}$ has the same properties as the general index $M(I)$ of (3.8), and simplifies to the expression in (3.10) when $L = 2$.

In a later paper, COHEN [22] introduced a modified form of kappa which allows for scaled disagreement or partial credit. Since certain patterns of disagreement between the observers may by more important than others, he proposed using a set of weights $\{w_{kk'}\}$ for $k, k' = 1, 2, \ldots, L$, which reflect the contribution of each cell in Table 6 to the measure of agreement. For example, in evaluating the reliability of a decision tree technique used to classify psychiatric patients into one of four groups, FELDMAN et al. [28] suggested weights which reflected the relative merits of each of the disagreements in diagnosis. In particular, they used weights which implied that it was less desirable to misclassify schizophrenia as excited or affective disorder than as a character disorder. Accordingly, in order to measure agreement with respect to a specified set of weights $\{w_{kk'}\}$, COHEN [22] defined a *weighted kappa* measure which is estimated by

$$\hat{\kappa}_w = \frac{p_o^* - p_e^*}{1 - p_e^*} \tag{4.6}$$

where

$$p_o^* = \sum_{k=1}^{L} \sum_{k'=1}^{L} w_{kk'} p_{kk'} \quad \text{and} \quad p_e^* = \sum_{k=1}^{L} \sum_{k'=1}^{L} w_{kk'} p_{k\cdot} p_{\cdot k'}. \tag{4.7}$$

In most cases, $0 \leqslant w_{kk'} \leqslant 1$ for all k, k', so that p_o^* is a weighted observed proportion of agreement, and p_e^* is the corresponding weighted proportion of agreement expected under the constraints of total independence. Furthermore, by choosing the weights in (4.8),

$$w_{kk'} = \begin{cases} 1, & k = k' \\ 0, & k \neq k', \end{cases} \tag{4.8}$$

$\hat{\kappa}_w$ in (4.6) simplifies to $\hat{\kappa}$ in (4.5).

EVERITT [26] derived the means and variances of both kappa and weighted kappa; but as indicated by FLEISS et al. [33], the standard errors given by COHEN [21, 22] and EVERITT [26] were derived under the assumptions of independence and fixed marginals, rather than under the single constraint of a fixed number of subjects, n. Accordingly, FLEISS et al. [33] calculated the unconditional large sample variance of weighted kappa as

$$\widehat{\operatorname{var}(\hat{\kappa}_w)} = \frac{1}{n(1-p_e^*)^4} \left\{ \sum_{k=1}^{L} \sum_{k'=1}^{L} p_{kk'} [w_{kk'}(1-p_e^*) - (\bar{w}_{k\cdot} + \bar{w}_{\cdot k'})(1-p_o^*)]^2 \right.$$
$$\left. - (p_o^* p_e^* - 2p_e^* + p_o^*)^2 \right\} \tag{4.9}$$

where

$$\bar{w}_{k\cdot} = \sum_{k'=1}^{L} w_{kk'} p_{\cdot k'} \quad \text{and} \quad \bar{w}_{\cdot k'} = \sum_{k=1}^{L} w_{kk'} p_{k\cdot}. \tag{4.10}$$

This expression in (4.9) reduces to the appropriate estimated variance of kappa in (4.5) by substituting the weights given in (4.8).

Agreement statistics such as weighted kappa in (4.6) can also be generated for ordinal scale data by assigning appropriate weights to each of the off-diagonal cells to reflect the degree of disagreement. One such selection of weights recommended by CICCHETTI [17, 18] is given by

$$w_{kk'} = 1 - \frac{|k-k'|}{(L-1)}. \tag{4.11}$$

Using the weights in (4.11), the Cicchetti test statistic for the significance of observer agreement is

$$z_C = \frac{p_o^* - p_e^*}{[\widehat{\text{var}}(p_o^*)]^{\frac{1}{2}}} \tag{4.12}$$

where

$$\widehat{\text{var}}(p_o^*) = \frac{1}{n-1}\left[\sum_{k=1}^{L}\sum_{k'=1}^{L} w_{kk'}^2 p_{kk'} - p_o^{*2}\right]. \tag{4.13}$$

Moreover, weights such as those in (4.11) can be used to generate the corresponding weighted kappa statistics in (4.6).

In situations where the L categories are not only ordinally scaled, but can be assumed to be equally spaced along some underlying continuum, discrete numerical integers such as $1, 2, \ldots, L$ can be assigned to the respective classes. In this context, by choosing the weights to be

$$w_{kk'} = 1-(k-k')^2, \tag{4.14}$$

COHEN [22] has shown that under observed marginal symmetry, weighted kappa in (4.6) is precisely equal to the product-moment correlation coefficient calculated on the integer-valued categories. Furthermore, FLEISS and COHEN [36] have shown that if the random effects model of (2.5) is assumed to hold for the data scored as $1, 2, \ldots, L$ by each of the two observers, the estimate of the intraclass correlation coefficient ϱ_2 in (2.9) is "asymptotically equal to" $\hat{\kappa}_w$ in (4.6) using the weights in (4.14).

Various other procedures involving the main diagonal of a square contingency table have been developed. For example, GOODMAN and KRUSKAL [45] proposed a measure of agreement of the type $M(I)$ in (3.8) based upon optimal prediction which can be estimated by

$$\lambda_r = \frac{\sum_{k=1}^{L} p_{kk} - \frac{1}{2}(p_{M\cdot} + p_{\cdot M})}{1-\frac{1}{2}(p_{M\cdot} + p_{\cdot M})}, \tag{4.15}$$

where $p_{M\cdot}$ and $p_{\cdot M}$ are the two marginal proportions corresponding to a hypothesized modal class. As defined here, λ_r ranges from -1 (when all the diagonal elements are zero and $p_{M\cdot}+p_{\cdot M}=1$) to $+1$ (when both observers are in complete agreement). In CHEN et al. [16], MANTEL and CRITTENDEN proposed a chi-square statistic with 1 d.f. as a test of agreement on the main diagonal cells. In another study reported by SPIERS and QUADE [87], the expected value for the (k, k')-th cell was considered to be a weighted average of the expected value under independence and the expected value with the diagonals inflated to the greatest possible extent. Using the method of minimum χ^2, estimates of these weights were derived, and then a test of independence was performed. LIGHT [66] has also recommended a chi-square statistic that is sensitive to the pattern of agreement on the main diagonal of the $L \times L$ table for two observers. Using the expected values based on independence and combining all the off-diagonal cells, his statistic χ_L^2 is given by

$$\chi_L^2 = n\left\{\sum_{k=1}^{L}\frac{(p_{kk}-p_{k\cdot}p_{\cdot k})^2}{p_{k\cdot}p_{\cdot k}} + \left[\sum_{k\neq k'}^{L}\sum^{L} p_{k\cdot}p_{\cdot k'}\right]^{-1}\left[\sum_{k\neq k'}^{L}\sum^{L}(p_{kk'}-p_{k\cdot}p_{\cdot k'})\right]^2\right\}, \quad (4.16)$$

which is asymptotically chi-square with L degrees of freedom under the hypothesis of independence. Here it can be noted that $\hat{\kappa}$ of (4.5) may be essentially zero, while χ_L^2 may be large and significantly different from zero. However, if χ_L^2 is near zero, $\hat{\kappa}$ will be necessarily near zero. As such, χ_L^2 reflects deviations from the expected pattern on the diagonal, while $\hat{\kappa}$ reflects the overall level of agreement.

4.3 *Measures of agreement among many observers*

Overall measures of inter-observer agreement have also been developed for the situation where each of $d > 2$ observers individually classify n subjects on an L-point scale. Moreover, most of these developments have been in terms of pairwise agreement considerations. For example, CARTWRIGHT [15] proposed an agreement coefficient which can be estimated by

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n}\left\{m_i \Big/ \binom{d}{2}\right\}, \quad (4.17)$$

where m_i is the number of pairs of raters in agreement on the classification of the i-th subject. In this context, $\hat{\alpha}$ ranges from 0 (no agreement) to $+1$ (perfect agreement) among the observers, since the $\{m_i\}$ range between 0 and $\binom{d}{2}$ for each subject. Thus, for dichotomous data, $\hat{\alpha}$ can be regarded as a complementary analogue to $\bar{D}$ in (3.19). Also, for nominal or ordinal scaled data, $\hat{\alpha}$ is identical to p_o in (4.3) when $d = 2$. In this respect, $\hat{\alpha}$ in (4.17) is essentially an uncorrected index of agreement, since the expected agreement calculated under such baseline constraints as total independence is not considered.

For this purpose, FLEISS [34] developed an extension of kappa in (4.5) for more

than two raters which does account for expected agreement under the baseline constraints of pairwise independence and marginal homogeneity. Specifically, he proposed an estimate of the overall observed proportion of agreement as

$$\bar{p}_o = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{L}\left\{\binom{u_{ik}}{2}\Big/\binom{d}{2}\right\}, \tag{4.18}$$

where u_{ik} is the number of assignments of the i-th subject to the k-th category by the d observers. In this regard, $\bar{p}_o$ is identically equal to $\hat{\alpha}$ in (4.17). Furthermore, his estimate of expected agreement can be written as

$$\bar{p}_e = \sum_{k=1}^{L} q_k^2, \tag{4.19}$$

where

$$q_k = \frac{1}{nd}\sum_{i=1}^{n} u_{ik} \tag{4.20}$$

is the overall proportion of assignments to the k-th category. Then in terms of (4.18) and (4.19) his extension of kappa to the case involving $d > 2$ observers can be estimated by

$$\hat{\kappa}_{(d)} = \frac{\bar{p}_o - \bar{p}_e}{1 - \bar{p}_e}. \tag{4.21}$$

Moreover, in certain situations it may be useful to partition an overall measure of agreement into component parts which reflect agreement for each of the L categories. For this purpose, FLEISS [34] showed that $\hat{\kappa}_{(d)}$ in (4.21) can be expressed alternatively as

$$\hat{\kappa}_{(d)} = \frac{\sum_{k=1}^{L} q_k(1-q_k)\hat{\kappa}_{(d)}^{(k)}}{\sum_{k=1}^{L} q_k(1-q_k)}, \tag{4.22}$$

where

$$\hat{\kappa}_{(d)}^{(k)} = \frac{\hat{Q}_k - q_k}{1 - q_k} \quad \text{for} \quad k = 1, 2, \ldots, L \tag{4.23}$$

are separate estimates of inter-observer agreement for each of the L categories. In this regard, $\hat{Q}_k$ in (4.23) is given by

$$\hat{Q}_k = \frac{\sum_{i=1}^{n}\left\{\binom{u_{ik}}{2}\Big/\binom{d}{2}\right\}}{nq_k}, \tag{4.24}$$

which is an estimate of the conditional probability of agreement between two

randomly selected observers on the assignment of a particular subject into the k-th category, given that the first observer classified the subject into the k-th category. As such, $\hat{\kappa}_{(d)}$ in (4.22) is shown to be a weighted average of the conditional agreement statistics in (4.23).

Similarly, LIGHT [66] has developed analogous statistics to (4.21) and (4.23) from a somewhat different point of view. In particular, these κ-type statistics involve pairwise agreement considerations based on fixed marginal distributions.

5 Measures of agreement with a standard

All the measures of reliability and the tests for marginal homogeneity discussed in the previous sections reflect the extent to which the observers agree among themselves on the classification of the same set of subjects. In this regard, since none of these measures of agreement involve comparisons with a "standard" or "correct" classification for each subject, they simply provide estimates for a type of internal consistency among the observers. Moreover, the absence of a known "true value" is precisely the motivation for using multiple observers in the measurement process.

Alternatively, in some situations, one of the observers may be the standard whose classification of the subjects is considered to be "correct". For example, in physical health exams the classifications by an expert diagnostician may be used as a standard for determining the validity of the diagnoses by a group of interns or medical students. Similarly, new clinical diagnostic procedures may be tested by comparing their results with those of a standard laboratory test as discussed in BENNETT [9]. Finally, "true values" may be created by using a separate panel of experts to provide a standard classification for each subject. Then the performance of other observers can be evaluated by comparing their classifications with the expert panel decisions.

The measurement of observer reliability within the context of a known standard involves issues which are quite different from those discussed previously. In particular, the major emphasis is given to comparisons between the observers and the standard, rather than to measures of agreement among the observers themselves. When the data are assumed to be quantitative, appropriate models can be readily obtained as special cases of the ANOVA models given previously. For example, by subtracting the "true values" from the observed values, the sources of variation for the resulting differences can be modeled by one of the ANOVA models given in Section 2. However, for the case where the data are categorial (*e.g.*, either dichotomous, nominal, or ordinal), such questions involving a standard have been evaluated by contingency table methods which have been developed from a somewhat different point of view from those discussed in Sections 3 and 4.

In particular, much work has been done for the dichotomous case involving a standard as reviewed recently by FEINSTEIN [27]. For this purpose, a variety of reliability and validity measures have been proposed in terms of the proportions from an appropriate 2×2 table. In this context, the n subjects are labeled as positive $(+)$ or negative $(-)$ by each of the measurement procedures, and then the resulting

Table 7. Classification by standard and test procedure on dichotomous scale

		test		
		+	−	total
standard	+	n_{11}	n_{12}	$n_{1\cdot}$
	−	n_{21}	n_{22}	$n_{2\cdot}$
	total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

data are cross-classified as shown in Table 7. YERUSHALMY [91] introduced the term *sensitivity* to denote the proportion of "true positives", ξ, and the term *specificity* to denote the proportion of "true negatives", η, associated with the test procedure. Using the notation in Table 7, the estimates of these quantities are given by

$$\hat{\xi} = \frac{n_{11}}{n_{1\cdot}} \qquad \hat{\eta} = \frac{n_{22}}{n_{2\cdot}}. \tag{5.1}$$

In both cases, these estimates range from 0 (for no agreement) to +1 (for perfect agreement) with the standard. Moreover, these statistics provide a separate estimate of agreement with the standard for the "positives" and for the "negatives". However, as indicated by FLEISS [37] and FEINSTEIN [27], if the test procedure is to be used for predictive purposes, such as a screening device, alternative reliability measures which indicate the positive accuracy, ϱ^+, and the negative accuracy, ϱ^-, are required. In this regard,

$$\hat{\varrho}^+ = \frac{n_{11}}{n_{\cdot 1}} \tag{5.2}$$

reflects the "true positive rate" of the test, and

$$\hat{\varrho}^- = \frac{n_{22}}{n_{\cdot 2}} \tag{5.3}$$

reflects the "true negative rate" of the test. In addition, these authors discussed the effect of the prevalence rate of the condition under investigation on the accuracy rates in (5.2) and (5.3). Otherwise, an overall estimate of validity for the test can be obtained from

$$p_0 = \frac{1}{n}(n_{11} + n_{22}), \tag{5.4}$$

which is directly analogous to the crude index of agreement in (3.3). Finally, BENNETT [9] showed how his results obtained in BENNETT [6, 7] could be used to compare the

sensitivity, specificity, and predictive value of several diagnostic procedures used on the same set of n subjects. These involve considerations directly analogous to the multiple observer case in Section 3.5.

More generally, when the classification scale is nominal or ordinal with $L > 2$ categories, the notions of sensitivity, specificity, and predictive value become much more complex to formulate, as discussed in FEINSTEIN [27]. However, each observer or test procedure can be compared with the standard in terms of a measure of validity directly analogous to (4.3) as an extension of (5.4). In addition to this, LIGHT [66] proposed a test of the joint agreement of the d observers with a standard in terms of an overall sum of the individual crude indices of agreement between each observer and the standard.

6 Some concluding remarks

Because the observer has been shown to be an important source of measurement error in data acquisition, reliability studies are conducted in experimental or survey situations to assess the extent of the observer variability. In all of these cases, the most common research design for a univariate response can be regarded as involving samples from s sub-populations of subjects on whom the response variable is measured separately by d observers. In this regard, observer reliability experiments or surveys involve research designs which produce repeated measurement data.

The questions of substantive interest in these repeated measurement situations are as follows:

1. Are there any differences among the sub-populations with respect to the distribution of the responses to the d observers?
2. Are there any differences among the distributions of responses to the d observers within each of the respective sub-populations?
3. Are there any differences among the sub-populations with respect to differences among the distributions of responses to the d observers? In other words, is there any observer x sub-population interaction?
4. Are there any differences among the sub-populations with respect to the overall agreement of the d observers on individual subjects?
5. Are there any differences in agreement among certain subsets of observers within each of the respective sub-populations?

As stated in KOCH et al. [63], questions (1)–(3) are directly analogous to the hypotheses of "no whole-plot effects", "no split-plot effects", and "no whole-plot by split-plot interaction" in standard split-plot experiments. In this context, question (1) addresses differences among the s sub-populations, question (2) involves the issue of inter-observer bias, and question (3) is concerned with the observer x sub-population interaction. In contrast to overall differences, questions (4)–(5) address the issue of agreement on a subject-to-subject basis. Here question (4) involves differences in measures of inter-observer agreement among the s sub-populations, and question (5)

is concerned with differences in pairwise agreement or agreement among subsets of the d observers.

When the data arising from observer reliability studies are quantitative measurements, tests for observer bias and measures of observer agreement are usually obtained from ANOVA models as discussed in Section 2. These models permit estimation of intraclass correlation coefficients for measures of agreement, and significance testing of the observer effects for the hypothesis of "no inter-observer bias". Even though assumptions of normality may not be warranted in certain cases, the ANOVA procedure discussed in ANDERSON and BANCROFT [1], SCHEFFÉ [81], and SEARLE [84], and the SSP procedure in KOCH [59, 60] still permit the estimation of the appropriate components of variance and the reliability coefficients used in assessing observer variability.

As reviewed in Sections 3 and 4, a wide variety of estimation and testing procedures have been developed to assess observer variability when the data are categorical. In these situations the response variable is classified into L nominal (or possibly ordinal) multinomial classes. Thus, the conceptual formulation of questions (1)–(5) may be undertaken in terms of an underlying $(s \times r)$ contingency table where $r = L^d$ represents the number of possible multivariate response profiles. Within this context, the first-order marginal distributions of response for each of the d observers contain most of the relevant information for dealing with questions (1)–(3). Furthermore, functions of the diagonal cells of various subtables provide the information for estimating and testing the significance of the agreement measures on a subject-to-subject basis. These quantities which reflect the extent to which the observers agree among themselves can be expressed as functions of the observed proportions obtained from the underlying contingency table. Accordingly, they can be analyzed within the scope of the general methodology for multivariate categorical data discussed in GRIZZLE et al. [47] (referred to as the GSK procedure). The GSK approach essentially involves a two-stage procedure:

i. the construction of the appropriate functions of the observed proportions which are directed at the relationships under investigation by a sequence of matrix formulations;
ii. the construction of test statistics for hypotheses involving these functions and the estimation of corresponding model parameters via weighted least squares computational algorithms.

For example, the GSK formulation for hypotheses of first-order marginal homogeneity in KOCH and REINFURT [61] and KOCH et al. [63] can be used for tests concerning inter-observer bias. Similarly, extensions of the GSK procedures discussed in FORTHOFER and KOCH [43] can be used to estimate and to model generalized kappa-type statistics for measures of inter-observer agreement. These topics and other applications of the GSK methodology to the multidimensional agreement problem are given in LANDIS [65].

Another approach to the analysis of multidimensional contingency tables is based

on maximum likelihood estimation within the framework of log-linear models as presented in BISHOP et al. [12]. Much of the research in this direction has been concerned with the analysis of multivariate relationships, and thus pertains to generalized measures of association. Otherwise, BISHOP et al. [12] have discussed agreement in two-way tables as a special case of association; and LIN [67] has applied maximum likelihood methods to more general multidimensional agreement problems.