Br. J. math. statist. Psychol. (1977), **30**, 98–103 *Printed in Great Britain*

Nominal scale response agreement as a generalized correlation

Lawrence Hubert

A further discussion of Brennan & Light's (1974) measure of nominal scale response agreement between two raters is given. Specifically, a monotonic function of Brennan & Light's statistic is obtained in terms of a generalized correlation coefficient and provided with a descriptive probabilistic interpretation. Under the different assumptions of fixed and variable marginal frequencies, simple numerical illustrations of the appropriate mean and variance formulae are included. The fixed marginal assumption was used by Brennan & Light, and thus our formulae in this case merely suggest convenient algebraic simplifications.

1. Introduction

In recent years the problem of measuring nominal scale response agreement between two judges has been approached in the psychological literature through a kappa statistic originally introduced by Cohen (see Cohen, 1960, 1968; Everitt, 1968; Fleiss *et al.* 1969; Light, 1971). Although the rationale justifying the kappa statistic has an obvious intuitive appeal for the particular application at hand, it is possible to define other indices that have interesting relationships to the popular measures of association introduced by Goodman & Kruskal (1954) and Kendall (1970). In particular, this paper further develops an alternative paradigm, introduced by Brennan & Light (1974), that depends on response *pairs* for each rater and fixed marginal frequencies.
In addition to some discussion of a related multinomial inference model with variable marginal frequencies, formulae that are somewhat simpler to use than those given by Brennan & Light are also provided.

2. Background

Suppose two raters assign n objects from a set S to a number of (nominal) classes. Formally, it is assumed that rater (1) uses R response categories $\{A_1, \ldots, A_R\}$, rater (2) uses C response categories $\{B_1, \ldots, B_C\}$ and n_{ij} is the number of objects from S placed in category A_i by rater (1) *and* in category B_j by rater (2). Alternatively, an $R \times C$ contingency table may be defined in the following manner, where standard notation is used for the various marginal sums:

		Rater (2)				
		B_1	B_2	$\cdots$	B_C	Sums
Rater (1)	A_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1.}$
	A_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	A_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	$n_{R.}$
	Sums	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.C}$	$n_{..} = n$

For the special case defined by $C = R$ and $A_i = B_i$, for all i, Cohen's agreement index κ is given by

$$\kappa = (P_0 - P_e)/(1-P_e),$$

where $P_0 = \sum_{i=1}^{R} n_{ii}/n$ is the observed proportion of 'agreement' and $P_e = \sum_{i=1}^{R} n_{i.}n_{.i}/n^2$ is the expected proportion of 'agreement' under the hypothesis of rater independence. A more general problem of interest in this paper allows the number of categories and their specification to differ from rater to rater. Clearly, the κ measure is inappropriate for this extension which permits each rater to use any partition of the objects in S as a final classification.

In discussing measures of response agreement in the general $R \times C$ contingency table, Brennan & Light (1974) suggested using the number of agreements in the categorization of all object pairs. In particular, there are four possible alternatives for each of the $\binom{n}{2}$ object pairs that may be formed from the elements in the set S (in the four alternatives below, the concepts of the 'same' and 'different' category refer to the responses of a single rater):

(i) objects from a pair are placed in the same category by rater (1) and in the same category by rater (2);

(ii) objects from a pair are placed in different categories by rater (1) and in different categories by rater (2);

(iii) objects from a pair are placed in different categories by rater (1) and in the same category by rater (2);

(iv) objects from a pair are placed in the same category by rater (1) and in different categories by rater (2).

An agreement in categorization for a specific object pair occurs when either condition (i) or (ii) obtains; alternatively, a disagreement occurs when either (iii) or (iv) obtains. If A denotes the number of pairs in (i) and (ii) and D denotes the number of pairs in (iii) and (iv), then from the results provided by Brennan & Light we have[†]

$$A = \binom{n}{2} + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\left(\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2\right)$$

and

$$A + D = \binom{n}{2}.$$

As a theoretical convenience in the discussion below, a particular monotonic function of A is considered that can be derived directly as a correlational index of agreement. Specifically, a scoring function for object pairs will be defined that is similar in form to the functions used in measuring monotonic relationships in contingency tables with ordered classes (Kendall, 1970).

3. A correlational index of agreement

Instead of categorizing the n objects from S in terms of an $R \times C$ contingency table, suppose the data are presented in the form of a sequence of n bivariate (non-numerical) observations:

$$(\mathbf{A}_1, \mathbf{B}_1), \ldots, (\mathbf{A}_n, \mathbf{B}_n),$$

[†] Various functions of A have been used for different purposes by other authors, e.g. Johnson (1968), Rand (1971) and Hartigan (1975).

where $(\mathbf{A}_k, \mathbf{B}_k) = (A_i, B_j)$ whenever rater (1) places the kth object from S into category A_i and rater (2) places the same object into category B_j. Furthermore, let $\mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n\}$, $\mathbf{B} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_n\}$, and define the following two functions on $\mathbf{A} \times \mathbf{A}$ and $\mathbf{B} \times \mathbf{B}$ corresponding to rater (1) and rater (2), respectively:

$$Q(\mathbf{A}_s, \mathbf{A}_t) = \begin{cases} 1 & \text{if } \mathbf{A}_s = \mathbf{A}_t, \\ -1 & \text{if } \mathbf{A}_s \neq \mathbf{A}_t, \end{cases} \qquad Q(\mathbf{B}_s, \mathbf{B}_t) = \begin{cases} 1 & \text{if } \mathbf{B}_s = \mathbf{B}_t, \\ -1 & \text{if } \mathbf{B}_s \neq \mathbf{B}_t, \end{cases}$$

for $1 \leqslant s \neq t \leqslant n$; also, it is assumed as a technical convenience that $Q(\mathbf{A}_s, \mathbf{A}_s) = Q(\mathbf{B}_s, \mathbf{B}_s) = 0$, $1 \leqslant s \leqslant n$. Given these preliminaries, Daniel's concept of a generalized correlation coefficient (see Kendall, 1970) can be extended by analogy to include a natural index of response agreement defined formally by

$$\Gamma = \left(\sum_{t=1}^{n}\sum_{s=1}^{n} Q(\mathbf{A}_s, \mathbf{A}_t) Q(\mathbf{B}_s, \mathbf{B}_t)\right) \Bigg/ \left(\sum_{t=1}^{n}\sum_{s=1}^{n} Q(\mathbf{A}_s, \mathbf{A}_t)^2 \sum_{t=1}^{n}\sum_{s=1}^{n} Q(\mathbf{B}_s, \mathbf{B}_t)^2\right)^{\frac{1}{2}}.$$

It is easily shown that

$$\Gamma = (A - D)/(A + D) = (A - D) \Bigg/ \binom{n}{2} = 1 + \left[2 \sum_{j=1}^{C}\sum_{i=1}^{R} n_{ij}^2 - \left(\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2\right)\right] \Bigg/ \binom{n}{2};$$

and that consequently, Γ is bounded between $+1$ and -1, with the extremes attained when $D = 0$ and $A = 0$, respectively. More significantly, Γ may be given a familiar type of operational definition considered desirable for a descriptive measure of association (Goodman & Kruskal, 1954):

> If two objects from S are chosen at random, Γ is the difference between the probability of selecting an 'agreement' pair minus the probability of selecting a 'disagreement' pair.

This probabilistic notion for Γ parallels very closely the well-known interpretation for the ordinal measure of association γ (see Goodman & Kruskal, 1954).

Although the main discussion in this paper deals with nominal or unordered categories $\{A_1, A_2, \ldots, A_R\}$ and $\{B_1, B_2, \ldots, B_C\}$, it is interesting to note that the same type of index Γ can be extended to include a number of alternative concepts of rater agreement. For instance, suppose the categories for both raters are ordered: $A_1 < A_2 < \ldots < A_R$; $B_1 < B_2 < \ldots < B_C$. If

$$Q'(\mathbf{A}_s, \mathbf{A}_t) = \begin{cases} +1 & \text{if } \mathbf{A}_s < \mathbf{A}_t, \\ -1 & \text{if } \mathbf{A}_s > \mathbf{A}_t, \\ 0 & \text{if } \mathbf{A}_s = \mathbf{A}_t, \end{cases} \qquad Q'(\mathbf{B}_s, \mathbf{B}_t) =)\begin{cases} +1 & \text{if } \mathbf{B}_s < \mathbf{B}_t \\ -1 & \text{if } \mathbf{B}_s > \mathbf{B}_t, \\ 0 & \text{if } \mathbf{B} \ = \mathbf{B}_t, \end{cases}$$

then the associated Γ statistic is Kendall's τ_b (Kendall, 1970). In an analogous fashion, various definitions for $Q(\cdot, \cdot)$ define Spearman's rank-order correlation and Pearson's product-moment correlation. For a more complete presentation of this paradigm, the reader is referred to Kendall (1970).

4. Mean and variance parameters using fixed marginals and rater independence

In suggesting A as a measure of nominal scale response agreement, Brennan & Light (1974) derive the exact mean and variance for A under the conditions of a hypergeometric model generating the $R \times C$ contingency table; or in other words, by assuming fixed marginals and 'independence' of the two raters. The variance formula they give is rather cumbersome to implement, but, fortunately, the same mean and variance parameters may be obtained using a slightly different inference model that also suggests a more compact expression for the variance. Specifically, it is assumed that the positions of the n given observations in **A** are fixed but all $n!$ possible permutations of the observations in **B** are equally likely under the hypothesis of rater independence. In terms of the $R \times C$ contingency table, the distribution of Γ over all $n!$ permutations is equivalent to what would be obtained from the well-known hypergeometric distribution, that is, using the standard distributional assumptions for a contingency table with fixed marginals. It should be noted that the same results will hold if the n given observations in **B** are fixed and all $n!$ possible permutations of the observations in **A** are considered equally likely.

Since a conditional randomization model is assumed to hold, only the non-constant portion of the index Γ, the numerator, has to be considered explicitly. In particular, let

$$\Lambda = \sum_{t=1}^{n}\sum_{s=1}^{n} Q(\mathbf{A}_s, \mathbf{A}_t)\,Q(\mathbf{B}_s, \mathbf{B}_t) = 2(A-D).$$

Then using the general results given by Mantel (1967) and simple algebra, we have

$$E(\Lambda) = [1/(n(n-1))]\,A_1 B_1,$$

$$\begin{aligned}\operatorname{var}(\Lambda) = 2n(n-1) - \{[1/(n(n-1))]A_1B_1\}^2 + [4/(n(n-1)(n-2))](A_2-A_3)(B_2-B_3)\\ + [1/(n(n-1)(n-2)(n-3))](A_1^2-4A_2+2A_3)(B_1^2-4B_2+2B_3),\end{aligned}$$

where

$$A_1 = 2\sum_{i=1}^{R} n_{i.}^2 - (n+1)n, \quad B_1 = 2\sum_{j=1}^{C} n_{.j}^2 - (n+1)n,$$

$$A_2 = 4\sum_{i=1}^{R} n_{i.}^3 - 4(n+1)\sum_{i=1}^{R} n_{i.}^2 + (n+1)^2 n, \quad B_2 = 4\sum_{j=1}^{C} n_{.j}^3 - 4(n+1)\sum_{j=1}^{C} n_{.j}^2 + (n+1)^2 n,$$

$$A_3 = B_3 = n(n-1).$$

In terms of the original index Γ,

$$E(\Gamma) = \left[1\Big/\binom{n}{2}\right]^2\left[\sum_{i=1}^{R} n_{i.}^2 - \binom{n+1}{2}\right]\left[\sum_{i=1}^{C} n_{.i}^2 - \binom{n+1}{2}\right],$$

$$\operatorname{var}(\Gamma) = \operatorname{var}(\Lambda)/(n(n-1))^2.$$

Furthermore, as a large sample approximation that ignores term of order $O(n^{-2})$, we obtain

$$\operatorname{var}(\Gamma) \approx (4/n)(\bar{A}_1^2 - \bar{A}_2)(\bar{B}_1^2 - \bar{B}_2),$$

where

$$\bar{A}_1 = 2\sum_{i=1}^{R}(n_{i.}/n)^2 - 1, \quad \bar{B}_1 = 2\sum_{j=1}^{C}(n_{.j}/n)^2 - 1,$$

$$\bar{A}_2 = 4\sum_{i=1}^{R}(n_{i.}/n)^3 - 4\sum_{i=1}^{R}(n_{i.}/n)^2 + 1, \quad \bar{B}_2 = 4\sum_{j=1}^{C}(n_{.j}/n)^3 - 4\sum_{j=1}^{C}(n_{.j}/n)^2 + 1.$$

As will be shown by a numerical example in a later section, the exact formulae given above for $E(\Gamma)$ and $\operatorname{var}(\Gamma)$ are simple transformations of the expressions given by Brennan & Light for $E(A)$ and $\operatorname{var}(A)$. Specifically, since $\Lambda = 4A - 2\binom{n}{2}$,

$$E(A) = \tfrac{1}{4}E(\Lambda) + n(n-1)/4 = (n(n-1)/4)(E(\Gamma)+1),$$

$$\operatorname{var}(A) = \tfrac{1}{16}\operatorname{var}(\Lambda) = \tfrac{1}{16}(n(n-1))^2\operatorname{var}(\Gamma).$$

5. Confidence intervals based on a multinomial model

If we assume that the $R \times C$ contingency table is generated from a single multinomial distribution with parameters $\{\pi_{ij}\}$, where $\sum_{j=1}^{C} \sum_{j=1}^{R} \pi_{ij} = 1$, then the general approach for constructing large sample confidence intervals discussed by Goodman & Kruskal (1972) may be followed. A population analogue for Γ, say γ, can be defined by replacing n_{ij}/n by π_{ij} and treating factors of $(n-1)$ as n (that is, ignoring the finite sampling correction used in the definition of Γ):

$$\gamma = 1 + 4\sum_{j=1}^{C} \sum_{i=1}^{R} \pi_{ij}^2 - 2\left(\sum_{i=1}^{R} \pi_{i.}^2 + \sum_{j=1}^{C} \pi_{.j}^2\right).$$

Then, using the general large sample variance formula for σ_γ^2 (see Brown, 1975), and after some minor algebraic simplification, we obtain

$$\hat{\sigma}_\gamma^2 = (2/n)^4\left[\sum_{j=1}^{C} \sum_{i=1}^{R} n_{ij}[2n_{ij} - (n_{i.} + n_{.j})]^2 - (1/n)\left(\sum_{j=1}^{C} \sum_{i=1}^{R} n_{ij}[2n_{ij} - (n_{i.} + n_{.j})]\right)^2\right].$$

In addition to providing an appropriate large sample variance for the construction of a large sample confidence interval on γ, the formula for $\hat{\sigma}_\gamma^2$ can be specialized to give a large sample variance statistic under the usual independence condition: $\pi_{ij} = \pi_{i.}\pi_{.j}$ for all i, j. For this latter application the appropriate variance term is found by replacing n_{ij} by $n_{i.}n_{.j}/n$ in $\hat{\sigma}_\gamma^2$:

$$\hat{\sigma}_\gamma^2 = (4/n)(\bar{A}_2\bar{B}_2 - \bar{A}_1^2\bar{B}_1^2),$$

where $\bar{A}_1$, $\bar{A}_2$, $\bar{B}_1$ and $\bar{B}_2$ are given in (1). It should be noted that the large sample variances based on the multinomial model and the fixed marginal model are different.

6. Numerical example

As an illustration of the formulae given in the previous sections, the same contingency table given by Brennan & Light (1974) is used. Table 1 presents the Brennan & Light data from two raters, each defining three classes in categorizing 15 objects.

Table 1. Brennan & Light's contingency table for two raters and 15 objects.

		Rater 2			
		B_1	B_2	B_3	
	A_1	4	0	1	5
Rater 1	A_2	1	1	3	5
	A_3	0	4	1	5
		5	5	5	15

By substituting into the various formulae listed previously, the following preliminary information can be found:

$$A = 75{\cdot}0, \quad D = 30{\cdot}0, \quad \Gamma = 0{\cdot}42857.$$

Using the general index Γ and the corresponding randomization model we obtain

$$E(\Gamma) = 0{\cdot}18367, \quad \text{var}\,(\Gamma) = 0{\cdot}007404, \quad Z = 2{\cdot}846,$$

where Z is a standardized value defined in the usual manner. These values may be used to immediately obtain the same numerical quantities calculated by Brennan & Light for A:

$$E(A) = 62{\cdot}143, \quad \operatorname{var}(A) = 20{\cdot}407, \quad Z = 2{\cdot}846.$$

The large sample variance under the general multinomial model is $\hat{\sigma}_\gamma^2 = 0{\cdot}030341$; thus, an approximate 95 per cent confidence interval on γ is given by the interval from 0·126 to 0·808, and is obtained by using $\hat{\gamma} \pm 1{\cdot}96\hat{\sigma}_\gamma$, where $\hat{\gamma}$ is now a maximum-likelihood estimate found by replacing π_{ij} by n_{ij}/n in the formula for γ, i.e. $\hat{\gamma} = 0{\cdot}467$. Under the independence hypothesis and as a direct consequence of equal row and column frequencies, the large sample variance value is zero. Although this last result is theoretically annoying, the asymptotic degeneracy of a variance expression under independence is not unusual in applications of this general type; in fact, numerous other illustrations of zero asymptotic variances are given by Goodman & Kruskal (1972) for a number of other measures of association. The large sample approximation for the fixed marginals model also suffers from this same asymptotic degeneracy.

Acknowledgement

Partial support for this research was supplied by the National Science Foundation through SOC-75-07870.

References

Brennan, R. L. & Light, R. J. (1974). Measuring agreement when two observers classify people into categories not defined in advance. *Br. J. math. statist. Psychol.* **27**, 154–163.

Brown, M. B. (1975). The asymptotic standard errors of some estimates of uncertainty in the two-way contingency table. *Psychometrika* **40**, 291–296.

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

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

Everitt, B. S. (1968). Moments of the statistics κ and weighted kappa. *Br. J. math. statist. Psychol.* **21**, 97–103.

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

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

Goodman, L. A. & Kruskal, W. H. (1972). Measures of association for cross classifications. IV: Simplification of asymptotic variances. *J. Am. statist. Ass.* **67**, 415–421.

Hartigan, J. A. (1975). *Clustering Algorithms.* New York: Wiley.

Johnson, S. C. (1968). Metric clustering. Unpublished manuscript.

Kendall, M. G. (1970). *Rank Correlation Methods.* London: Griffin.

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

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. statist. Ass.* **66**, 846–850.

Received 11 February 1976; revised version received 13 October 1976.

Requests for reprints should be addressed to Lawrence Hubert, Department of Educational Psychology, The University of Wisconsin, 1025 West Johnson Street, Madison, Wisconsin 53706, USA.