ASSESSING PAIRWISE INTERRATER AGREEMENT FOR MULTIPLE ATTRIBUTE RESPONSES: THE "NO ATTRIBUTE" RESPONSE SITUATION

by

Lawrence L. Kupper and Kerry B. Hafner

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1845

February 1988

ASSESSING PAIRWISE INTERRATER AGREEMENT FOR MULTIPLE ATTRIBUTE RESPONSES: THE "NO ATTRIBUTE" RESPONSE SITUATION

Lawrence L. Kupper and Kerry B. Hafner
Department of Biostatistics, School of Public Health,
University of North Carolina at Chapel Hill,
Chapel Hill, North Carolina 27599-7400, U.S.A.

SUMMARY

Kupper and Hafner (1988) recently developed methods for assessing the extent of interrater agreement when each observational unit is characterized by *one or more* nominal attributes. They proposed a new two-rater concordance statistic, and provided associated inference-making tools. Except for the special case when each rater chooses *exactly one attribute* to describe each unit, their methodology does not permit a rater to claim that a unit has none of the attributes under consideration. This paper closes that methodologic gap. A numerical example is also provided.

Key Words: Chance-corrected agreement statistics; Non-central hypergeometric distribution; Kappa statistic; Conditional distribution theory; Nominal attributes.

1. Introduction and Notation

Consider a study in which two similarly trained raters, A and B, independently evaluate each of n units (*e.g.*, X-rays, human or animal subjects, etc.). From a prespecified group of distinct nominal attributes, each rater is to assign to each of the n units either *none* of these attributes or a subset containing *at least one* of them. The purpose of this study is to quantify the extent of agreement between raters A and B, and then to make appropriate statistical inferences.

Kupper and Hafner (1988) have recently developed methods for assessing pairwise interrater agreement for multiple attribute response data. However, except in the special case when each rater is to choose exactly one nominal attribute to describe each unit, their methods cannot be applied to the situation described above. In this paper, we extend the Kupper-Hafner methodology to deal with multiple attribute response data when at least one of the two raters has assigned *none* of the specified attributes to one or more units.

Following Kupper and Hafner (1988), let $\mathfrak{K}$ denote a prespecified set of k (≥ 3) elements. In this set, let $(k-1)$ of the elements denote distinct nominal attributes, specified so that the choice of any one attribute does not preclude the choice of any other attribute; the k-th element in the set $\mathfrak{K}$ represents the choice of *none* of the other $(k-1)$ attributes.

For the i-th unit, $i=1, 2,\ldots, n$, let the random variables A_i and B_i denote the numbers of elements chosen from the set $\mathfrak{K}$ by raters A and B, respectively. For the set $\mathfrak{K}$ as defined above, it is impossible for either rater to choose all k elements to describe the i-th unit; more specifically, $1 \leq A_i \leq (k-1)$ and $1 \leq B_i \leq (k-1)$. Given $A_i=a_i$ and $B_i=b_i$, consider the subsets $\mathcal{A}_i$ and $\mathcal{B}_i$, where $\mathcal{A}_i$ is the subset of elements from $\mathfrak{K}$ chosen by rater A to describe the i-th unit, and where $\mathcal{B}_i$ is defined analogously for rater B. Hence, $\text{Card}(\mathcal{A}_i)=a_i$ and $\text{Card}(\mathcal{B}_i)=b_i$.

Define the random variable X_i to be the number of elements common to the sets $\mathcal{A}_i$ and $\mathcal{B}_i$, namely,

$$X_i = \text{Card}(\mathcal{A}_i \cap \mathcal{B}_i).$$

Kupper and Hafner (1988) defined the "observed proportion of concordance" between raters A and B for the i-th unit to be

$$\hat{\pi}_i = X_i/\max(a_i, b_i), \quad i=1, 2,\ldots, n. \tag{1}$$

When $\mathcal{A}_i=\mathcal{B}_i$, indicating perfect concordance between raters A and B for the i-th unit, then $\hat{\pi}_i=1$. At the other extreme, when $\mathcal{A}_i \cap \mathcal{B}_i=\emptyset$, reflecting complete discord between raters A and B, then $\hat{\pi}_i=0$. The *overall* observed concordance between raters A and B can then be defined as

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_i. \tag{2}$$

Since some of this observed agreement between raters A and B may be due to chance alone, the crude statistic (2) should be corrected in an appropriate manner (for related discussion, see Fleiss, 1975); methods for making such a correction are provided in the next section.

2. Derivation of the Chance-Corrected Concordance Statistic $\hat{C}^*_{AB}$

All of the developments to follow are conditional on $A_i=a_i$ and $B_i=b_i$, $i=1, 2,..., n$. To simplify the notation, let $m_i=\min(a_i, b_i)$ and $M_i=\max(a_i, b_i)$; in our situation, $1 \le m_i \le M_i \le (k-1)$.

To correct expression (2) for chance agreement, we need to determine the conditional null distribution of X_i given that $A_i=a_i$ and $B_i=b_i$. If both raters choose their subsets $\mathcal{A}_i$ and $\mathcal{B}_i$ to describe the i-th unit in a completely random manner, then it can be shown (with some effort) that the null distribution of X_i is

$$pr_o(X_i=x_i)=\frac{\binom{a_i}{x_i}\binom{k-a_i-\delta_i}{b_i-x_i}}{\binom{k-\delta_i}{b_i}}, \quad i=1, 2,..., n ; \tag{3}$$

here, $\max(0, a_i+b_i-k+\delta_i) \le x_i \le m_i$, and the indicator variable $\delta_i=1$ if $m_i>1$, and $\delta_i=0$ if $m_i=1$. Hence, from (1) and (3), the null mean and variance of $\hat{\pi}_i$ are, respectively,

$$E_o(\hat{\pi}_i)=\frac{m_i}{(k-\delta_i)}, \tag{4}$$

and

$$V_o(\hat{\pi}_i)=\frac{(k-a_i-\delta_i)(k-b_i-\delta_i)m_i}{(k-\delta_i)^2(k-\delta_i-1)M_i}. \tag{5}$$

It follows from (2) and (4) that the null mean of $\hat{\pi}$ is

$$\pi^*_o = E_o(\hat{\pi})=\frac{\sum_{i=1}^{n}[m_i-\frac{1}{k}(1-\delta_i)]}{n(k-1)}. \tag{6}$$

Finally, for multiple attribute response data where both raters have independently assigned to each of n units either none or at least one of the $(k-1)$ attributes, our proposed overall chance-corrected measure of concordance between raters A and B is

$$\hat{C}^{*}_{AB} = \frac{\hat{\pi} - \pi^{*}_{o}}{1 - \pi^{*}_{o}} . \tag{7}$$

Under the null hypothesis that both raters A and B choose their subsets of attributes completely at random, $E_o(\hat{C}^{*}_{AB}) = 0$. When $\hat{\pi} = 1$, there is perfect concordance between raters A and B, and so $\hat{C}^{*}_{AB} = 1$. The minimum value of $\hat{C}^{*}_{AB}$ is $-\pi^{*}_{o}/(1-\pi^{*}_{o})$, which occurs when $\hat{\pi} = 0$; this minimum can never be smaller than -1.

3. Inference Methods

Once the sample concordance statistic $\hat{C}^{*}_{AB}$ has been calculated, it is appropriate to use it to test hypotheses about, and to construct confidence intervals for, the population concordance $C^{*}_{AB} = E(\hat{C}^{*}_{AB})$. For example, a test of H_0: $C^{*}_{AB} = 0$ versus H_1: $C^{*}_{AB} > 0$ (*i.e.*, a test to assess whether the observed concordance is significantly greater than that due to chance alone) can be based on the statistic

$$Z = \frac{\hat{C}^{*}_{AB}}{\sqrt{V_o(\hat{C}^{*}_{AB})}} , \tag{8}$$

where, from (5) and (7),

$$V_o(\hat{C}^{*}_{AB}) = [n(1-\pi^{*}_{o})]^{-2} \sum_{i=1}^{n} V_o(\hat{\pi}_i) .$$

Under H_0, the statistic (8) is approximately standard normal for large n.

To develop an approximate $100(1-\alpha)\%$ confidence interval for C^{*}_{AB}, it is necessary to find an estimate for the non-null variance $V(\hat{C}^{*}_{AB})$ of $\hat{C}^{*}_{AB}$. As Kupper and Hafner (1988) have shown, the non-central hypergeometric distribution (see Cox, 1970) is applicable here. In particular, given $A_i = a_i$ and $B_i = b_i$, it can be shown that the appropriate conditional non-central distribution of X_i is

$$\text{pr}(X_i=x_i;\,\psi_i)=\frac{\binom{a_i}{x_i}\binom{k-a_i-\delta_i}{b_i-x_i}\psi_i^{x_i}}{\sum_{u_i}\binom{a_i}{u_i}\binom{k-a_i-\delta_i}{b_i-u_i}\psi_i^{u_i}}\,, \tag{9}$$

where the summation in the denominator is over all integer values of u_i satisfying $\max(0,\, a_i+b_i-k+\delta_i)\le u_i\le m_i$. When the non-centrality parameter $\psi_i=1$, expression (9) reduces to the null distribution of X_i given by (3).

Under the assumption that $\psi_i=\psi$, $i=1, 2,\ldots, n$, the appropriate estimate $\hat{\psi}$ of ψ is obtained as the solution to the equation

$$\sum_{i=1}^{n} x_i=\sum_{i=1}^{n} E(X_i;\,\psi)\,, \tag{10}$$

where $E(X_i;\,\psi)$ is obtained using (9). Breslow (1981) has shown that a good approximation to $\hat{\psi}$ obtained via (10) is the Mantel-Haenszel estimate

$$\hat{\psi}_{mh}=\frac{\sum_{i=1}^{n} x_i\,(k-a_i-b_i+x_i-\delta_i)}{\sum_{i=1}^{n}(a_i-x_i)\,(b_i-x_i)}\,.$$

Thus, an approximate large-sample confidence interval for C^*_{AB} can be computed as

$$\hat{C}^*_{AB}\pm Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{C}^*_{AB})}\,,$$

in which

$$\hat{V}(\hat{C}^*_{AB})=[n(1-\pi^*_o)]^{-2}\sum_{i=1}^{n} V(X_i;\,\hat{\psi}_{mh})/M_i^2\,,$$

and where $V(X_i;\,\hat{\psi}_{mh})$ is obtained from (9) with $\hat{\psi}_{mh}$ substituted for ψ_i.

3. A Numerical Example

Dental X-ray data from a study conducted by Kantor, Reiskin, and Lurie (1985) are used in this example to illustrate the general concordance assessment methods proposed in this paper. The purpose of their study was to compare the X-ray quality of a standard film ("U") to a faster speed film ("E") that reduces patient X-ray exposure by approximately 50 percent. See Kupper and Hafner (1988) for details concerning an analysis of a subset of these data.

To compare the two film types, each of three raters (A, B, and C) independently evaluated n=106 X-ray films (51 of type "U" and 55 of type "E") to determine which, *if any*, of the k=14 teeth on each film had carious lesions (cavities). Note that the Kupper-Hafner (1988) subset analysis did *not* consider those X-ray films for which at least one of the three raters found no cavities.

The frequencies of the numbers of cavities reported by the three raters are summarized in Table 1. Note that the rater distributions by film type are fairly similar. However, the extent of agreement among the raters cannot be assessed using the data as displayed in Table 1; further information (which can be obtained from the authors upon request) is required to address this issue.

Insert Table 1 here.

The methods developed in this paper are used to assess the extent of agreement between each pair of raters. The three pairwise concordance statistics ($\hat{C}^*_{AB}$, $\hat{C}^*_{AC}$, and $\hat{C}^*_{BC}$), Z-tests of H_0: $C^*=0$ versus H_1: $C^*>0$, and approximate 95% confidence intervals for each film speed type are displayed in Table 2. The pairwise concordance statistics $\hat{C}^*_{AC}$ and $\hat{C}^*_{BC}$ for the "E" film are larger than those for the "U" film, but the opposite holds for $\hat{C}^*_{AB}$. Also, all of the concordance statistics are significantly greater than what would be expected by chance alone.

Insert Table 2 here.

To assess whether the expected concordances for the two film speed types are equal or not (*i.e.*, to test H_0: $C^{*U} - C^{*E} = 0$ versus H_1: $C^{*U} - C^{*E} \neq 0$), a test statistic of the form

$$Z = \frac{\hat{C}^{*U} - \hat{C}^{*E}}{\sqrt{\hat{V}(\hat{C}^{*U}) + \hat{V}(\hat{C}^{*E})}}$$

can be used, where the superscripts denote the two film speeds. For large n, this test statistic will have an approximate standard normal distribution if the null hypothesis is true. In addition, an approximate $100(1-\alpha)\%$ confidence interval for the true concordance difference $(C^{*U} - C^{*E})$ is

$$(\hat{C}^{*U} - \hat{C}^{*E}) \pm Z_{1-\alpha/2} \sqrt{\hat{V}(\hat{C}^{*U}) + \hat{V}(\hat{C}^{*E})} \, .$$

The results of this comparative analysis for the two film speed types are displayed in Table 3. Based on these findings, the images on the two different speed films appear to be equally readable by the three raters.

Insert Table 3 here.

4. Remarks

It is of interest to discuss briefly how the methodology in this paper relates to that developed earlier by Kupper and Hafner (1988).

The two methodologies are equivalent when each rater is to choose *exactly one* element from the set $\mathfrak{X}$ to describe each unit (*i.e.*, when $m_i = M_i = 1$ for all i). In that special case, the k elements of $\mathfrak{X}$ can correspond either to k distinct nominal attributes, or to $(k-1)$ such attributes and a k-th element representing the choice of none of these $(k-1)$ attributes. This special situation was discussed in detail by Kupper and Hafner (1988).

If the observed data are such that $\delta_i = 1$ for all i, then the raters are *always* selecting from the set of $(k-1)$ attributes to describe each and every unit, even though they had the option *not* to do so (*i.e.*, they could have selected the k-th element of $\mathfrak{K}$, namely, the element corresponding to none of these attributes). In this extremely rare instance, the results here specialize to those given by Kupper and Hafner (1988) when k is replaced by $(k+1)$ in all formulas in this paper.

Despite these two very special cases, giving raters the freedom *not* to choose any of the $(k-1)$ specified nominal attributes necessitates the use of alternative probabilistic arguments. Thus, for all other situations in which "no attribute" rater responses are plausible, the procedures developed in this paper are appropriate and are non-trivial extensions of the Kupper and Hafner (1988) methods.

Acknowledgements

The authors would like to thank Dr. Mel Kantor, DDS for providing the dental data used in the numerical example. The second author was partially funded by N.I.E.H.S. training grant #2 T32 ES07018 while a Ph.D. student at the University of North Carolina in Chapel Hill.

Table 1. Frequency distributions of the numbers of cavities reported by each rater for each film speed type.

	Number of cavities for "U" film type							
Rater	0	1	2	3	4	5	6	7
A	21	15	4	3	5	2	1	0
B	25	13	6	5	0	2	0	0
C	14	7	9	10	6	2	2	1

	Number of cavities for "E" film type							
Rater	0	1	2	3	4	5	6	7
A	22	12	10	5	4	1	1	0
B	27	15	4	4	4	0	1	0
C	18	8	13	6	7	1	1	1

Table 2. Pairwise concordance statistics and associated inferential statistics for each film speed type.

Speed = "U"

Raters	$\hat{C}^*$	$V_o(\hat{C}^*)$	Z	$\hat{\psi}_{mh}$	$\hat{V}(\hat{C}^*)$	Approximate 95% C.I.
AB	0.591	0.0012	12.4	38.88	0.0029	(0.484, 0.697)
AC	0.480	0.0010	11.6	19.17	0.0026	(0.379, 0.580)
BC	0.421	0.0010	11.9	14.78	0.0026	(0.320, 0.521)

Speed = "E"

Raters	$\hat{C}^*$	$V_o(\hat{C}^*)$	Z	$\hat{\psi}_{mh}$	$\hat{V}(\hat{C}^*)$	Approximate 95% C.I.
AB	0.536	0.0011	10.8	21.06	0.0029	(0.430, 0.643)
AC	0.562	0.0010	10.9	18.77	0.0025	(0.464, 0.659)
BC	0.495	0.0010	11.9	17.70	0.0026	(0.394, 0.596)

Table 3. Inferences comparing film speed types.

Raters	Z	Approximate 95% C.I.
AB	0.72	(−0.094, 0.204)
AC	1.15	(−0.058, 0.222)
BC	1.03	(−0.067, 0.215)

REFERENCES

Breslow, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika* 68, 73-84.

Cox, D.R. (1970). *The Analysis of Binary Data.* London: Methuen.

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

Kantor, M.L., Reiskin, A.B., and Lurie, A.G. (1985). A clinical comparison of X-ray films for detection of proximal surface caries. *Journal of the American Dental Association* 111, 967-969.

Kupper, L.L. and Hafner, K.B. (1988). The assessment of interrater agreement for multiple attribute responses. University of North Carolina Institute of Statistics Mimeo Series No. 1843 (January).