Communications of the Korean Statistical Society
2009, Vol. 16, No. 6, 1023–1029

Comparing More than Two Agreement Measures Using Marginal Association

Myongsik Oh[1,a]

[a]Department of Statistics, Pusan University of Foreign Studies

Abstract

Oh (2009) has proposed a likelihood ratio test for comparing two agreements for dependent observations based on the concept of marginal homogeneity and marginal stochastic ordering. In this paper we consider the comparison of more than two agreement measures. Simple ordering and simple tree ordering among agreement measures are investigated. Some test procedures, including likelihood ratio test, are discussed.

Keywords: Agreement, chi-bar-square, marginal homogeneity, simple order, simple tree order, stochastic ordering.

1. Introduction

Suppose more than two groups of raters classify a sample of subjects using the same ordered categorical scale and we want to compare these agreements. If there are only two observers in each group we may compute kappa statistics for each group and compare them. Fleiss and Cohen (1973) studied the measuring agreement for ordinal categorical data using so-called weighted kappa defined as

$$\kappa_w = \frac{\sum\sum w_{ij}\pi_{ij} - \sum\sum w_{ij}\pi_{i+}\pi_{+j}}{1 - \sum\sum w_{ij}\pi_{i+}\pi_{+j}}, \tag{1.1}$$

where the weights are chosen to be $w_{ij} = 1 - (i-j)^2/(I-1)^2$ and $i, j = 1, 2, \ldots, I$.

However, these kappas are apparently not independent and hence no usual k-sample inference tools can be directly applied. McKenzie *et al.* (1996) studied the comparison of two kappa statistics obtained from each of two dependent 2 by 2 tables. Donner *et al.* (2000) studied the test for the equality of two dependent kappa statistics. The two observers in each of the two groups use a binary scale to rate the subjects. This research, however, can be applied only when a binary scale is used for rating subjects.

Recently Oh (2008, 2009) has proposed test procedures including likelihood ratio test based on the marginal association for the two-sample case. In this paper we are going to extend Oh (2009)'s result to the case of multi-sample. This extension seems to be quite straightforward but need some intensive computational works as well as some statistical issues arising in order restricted statistical inference.

[1] Professor, Department of Statistics, Pusan University of Foreign Studies, 15 SukPo-Ro, Nam-Gu, Pusan 608-738, Korea.
E-mail: moh@pufs.ac.kr

2. Comparison of g Independent Agreements

Oh (2009) has proposed an alternative way of comparing agreements which does not require the assumption of equal marginal probabilities. In this section we assume that there are only two rates in each group and hence we can form a square contingency table. In a square contingency table which displays the joint rating of two observers, the diagonal cells are the most strongly agreed and the lower left and upper right cells are the least strongly agreed. The closer to the diagonal, the stronger the agreement. Based on this concept we define agreement ordering as follows. Let π_{ij} and π'_{ij} be cell probabilities of two square contingency tables with the same dimension. Then we say that the contingency table corresponding to $\boldsymbol{\pi}'$ has stronger agreement than the contingency table corresponding to $\boldsymbol{\pi}$ if

$$\sum_{i=1}^{k}\pi'_{ii} \geq \sum_{i=1}^{k}\pi_{ii},$$
$$\sum_{i=1}^{k}\pi'_{ii} + \sum_{r=1}^{\ell}\left\{\sum_{i=1}^{k-r}(\pi'_{i,i+r} + \pi'_{i+r,i})\right\} \geq \sum_{i=1}^{k}\pi_{ii} + \sum_{r=1}^{\ell}\left\{\sum_{i=1}^{k-r}(\pi_{i,i+r} + \pi_{i+r,i})\right\} \tag{2.1}$$

for $\ell = 1, \ldots, k-1$. We denote this $\boldsymbol{\pi} \preceq \boldsymbol{\pi}^{'}$.

Since this agreement ordering (2.1) is closely related to a stochastic ordering the statistical inference concerning agreement ordering follows that of stochastic ordering between two multinomial parameters. See Robertson and Wright (1981) and Oh (2009).

Now consider g square contingency tables labeled $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_g$ and impose a partial order induced by (2.1) among g tables. First we consider simple ordering. Suppose

$$\boldsymbol{\pi}_1 \preceq \boldsymbol{\pi}_2 \preceq \cdots \preceq \boldsymbol{\pi}_g.$$

Then this is going to be a problem of g stochastically ordered distributions. There are vast literature concerning these problems. Dardanoni and Forcina (1998), Feng and Wang (2007), El Barmi and Johnson (2006), El Barmi and Mukerjee (2005), Wang (1996) are among others.

Next we consider simple tree ordering. For example we may consider the following restriction on tables.

$$[\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_{k-1}] \preceq \boldsymbol{\pi}_g \quad \text{or} \quad [\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_{g-1}] \succeq \boldsymbol{\pi}_g.$$

To our best knowledge, no statistical inference procedures concerning this problem has been studied. We leave it as a future study.

3. Comparison of g Dependent Agreements

Suppose g groups of observers classify a sample of subjects using the same ordered categorical scale and that we are interested in comparing g agreements. If there are only two observers in each groups, then we may compute g kappas and compare these these g measurements. However, these measurements are apparently not independent and hence no usual g-sample inference tool nor the test procedure proposed in Section 2 can be directly applied. Donner *et al.* (2000), McKenzie *et al.* (1996) studied the comparison of two dependent (correlated) kappas. However, the categorical scales used in their studies were restricted to binary. Now we are going to consider this problem of arbitrary scales.

Let $\mathbf{X}_i = (\mathbf{X}_{1i}, \mathbf{X}_{2i}, \ldots, \mathbf{X}_{gi})$ be the score vector of i^{th} subject given by the raters, where $\mathbf{X}_{\ell i} = (X_{\ell 1i}, X_{\ell 2i}, \ldots, X_{\ell k_\ell i})$, $j = 1, 2, \ldots, k_\ell$, $\ell = 1, 2, \ldots, g$ and $X_{\ell ji}$ be the score of i^{th} subject given by the

j^{th} rater in ℓ^{th} group of raters. Consider g positive valued functions $D_1, D_2, \ldots, D_g$ such that

(1) $D_\ell : \mathbf{R}^{k_\ell} \to \mathbf{R}^+$, for $\ell = 1, 2, \ldots, g$,

(2) $D_\ell(\mathbf{x}) = D_\ell(\mathbf{x}_\gamma)$ for any permutation $\mathbf{x}_\gamma$ of $\mathbf{x}$,

(3) $D_\ell((a, a, \ldots, a)) = 0$ for all $a > 0$,

(4) $D_\ell(\cdot)$ is convex function over $\mathbf{R}^{k_\ell}$,

(5) $\{y_i \in \mathbf{R}^+ : y_i = D_1(X_{1i})\} = \{y_i \in \mathbf{R}^+ : y_i = D_2(X_{2i})\} = \cdots = \{y_i \in \mathbf{R}^+ : y_i = D_g(X_{gi})\}$.

Note that D_ℓ can be served as a measure of dispersion. For details, see Gilula and Haberman (1995). Then we can use this function as a measure of agreement of raters within subject. It is said to be more agreed for the smaller value of D_ℓ. For example, D_ℓ can be chosen to be

$$D_\ell(\mathbf{X}_{\ell i}) = \max_{j=1,\ldots,k_\ell} \{X_{\ell j i}\} - \min_{j=1,\ldots,k_\ell} \{X_{\ell j i}\}.$$

Now suppose, for all $t > 0$,

$$\frac{\#\{i : D_{\ell_1}(\mathbf{X}_{\ell_1 i}) \le t,\ i = 1, \ldots, n\}}{n} \ge \frac{\#\{i : D_{\ell_2}(\mathbf{X}_{\ell_2 i}) \le t,\ i = 1, \ldots, n\}}{n}.$$

This means that the proportion of more agreed subjects for ℓ_1^{th} observer group is always greater than ℓ_2^{th} observer group. As we have seen in Oh (2009), this is clearly related to stochastic ordering.

Assume that the set of possible values of D_ℓ's be $\{s_1, \ldots, s_k\}$. Let

$$p_{i_1 i_2 \ldots i_g} = \frac{\#\{i : D_\ell(\mathbf{X}_{\ell i}) = s_{i_\ell},\ \ell = 1, \ldots, g,\ i = 1, \ldots, n\}}{n},$$

for $i_\ell \in \{1, \ldots, k\}$ where $\ell = 1, 2 \ldots, g$. Also let $\hat{p}_{i_1 i_2 \ldots i_g}$ be the observation of $p_{i_1 i_2 \ldots i_g}$ and $n_{i_1 i_2 \ldots i_g} = n\hat{p}_{i_1 i_2 \ldots i_g}$. Note that $p_{i_1 i_2 \ldots i_g}$ form a g-dimensional square contingency table. Define marginal probability $\mathbf{p}^{(\ell)} = (p_1^{(\ell)}, p_2^{(\ell)}, \ldots, p_k^{(\ell)})$ as follows;

$$p_j^{(\ell)} = \sum_{\substack{i_\ell = j \\ i_h = 1, \ldots, k \text{ if } h \neq \ell}} p_{i_1 i_2 \ldots i_g}, \quad j = 1, 2, \ldots, k,\ \ell = 1, 2, \ldots, g.$$

Consider the hypothesis

$$H_0 : \mathbf{p}^{(1)} = \mathbf{p}^{(2)} = \cdots = \mathbf{p}^{(g)}, \tag{3.1}$$

where

$$\mathbf{p}^{(i)} = \mathbf{p}^{(j)} \text{ if and only if } p_h^{(i)} = p_h^{(j)}, \quad \text{for } h = 1, \ldots, k.$$

This is so-called marginal homogeneity of g-dimensional square contingency table.

Next we consider the restricted alternative hypotheses. Let $I = \{1, 2, \ldots, g\}$ be an index set and $\preceq$ be a partial order on I. Here we consider an alternative hypothesis related to this partial ordering. Suppose $i \preceq j$ implies $\mathbf{p}^{(i)} \le_S \mathbf{p}^{(j)}$, where

$$\mathbf{p}^{(i)} \le_S \mathbf{p}^{(j)} \text{ if and only if } \sum_{\ell=1}^{h} p_\ell^{(i)} \le \sum_{\ell=1}^{h} p_\ell^{(j)}, \quad \text{for } h = 1, \ldots, k. \tag{3.2}$$

If (3.2) holds for each pair (i, j) such that $i \preceq j,\ i, j \in I$ then we say that $\mathbf{p}^{(\ell)}$'s is isotonic with respect to $\preceq$ on I.

Now we consider the following alternative hypotheses

$$H_a : p^{(\ell)}\text{'s is isotonic with respect to } \preceq \text{ on } I.$$

There are many types of partial ordering we may interested in. Simple ordering and simple tree ordering are, however, two most widely used partial orderings. In this paper we consider these two partial ordering only since these are the indicative of all other orderings. First we consider simple ordering.

3.1. Simple order

Consider the following hypothesis.

$$H_{a_1} : \mathbf{p}^{(1)} \le_S \mathbf{p}^{(2)} \le_S \cdots \le_S \mathbf{p}^{(g)}. \tag{3.3}$$

This is a simple ordering or an increasing order.

Let $\mathbf{C} = \{c_{ij}\}_{k^2\times k}$, $i = 1,\ldots,k^2,\ j = 1,\ldots,k$, where

$$c_{ij} = I_{\{k\times(j-1)+1,\ldots,k\times j\}}(i) - I_{\{mod(i-1,k)+1\}}(j),$$

where $I_A(\cdot)$ is an indicator function, and mod(i,k) is residue when i is divided by k. Let $\mathbf{p} = (\{p_{i_1 i_2 \ldots i_g}, i_1, i_2 \ldots i_g = 1, 2, \ldots, k\})'$. Let

$$\mathbf{A} = [A_1, \ldots, A_{g-1}],$$

where

$$A_i = \underbrace{\mathbf{1}\otimes\cdots\otimes\mathbf{1}}_{i-1}\otimes\mathbf{C}\otimes\underbrace{\mathbf{1}\otimes\cdots\otimes\mathbf{1}}_{g-i-1},$$

and $\mathbf{1}_{k\times 1} = (1, 1, \ldots, 1)'$ and $\bigotimes$ is kronecker's product.

Then restrictions (3.1) and (3.3) are re-expressed by

$$\mathbf{A}'\mathbf{p} = \mathbf{0}, \qquad \text{and} \tag{3.4}$$

$$\mathbf{A}'\mathbf{p} \le \mathbf{0}, \tag{3.5}$$

respectively, where the equality and inequality between vectors are componentwise. Let

$$\mathbf{B} = \mathbf{I}_{g-1} \otimes B$$

where $B = \{b_{ij}\}_{k\times(k-1)}$, $i = 1, \ldots, k,\ j = 1, \ldots, k-1$ with $b_{ij} = 1$ if $i \le j$, $= 0$, otherwise, and $\mathbf{I}_{g-1}$ is identity matrix with dimension $g-1$. To eliminate redundancy in (3.4) and (3.5), we multiply by B and have

$$\mathbf{B}'\mathbf{A}'\mathbf{p} = \mathbf{0}, \qquad \text{and} \tag{3.6}$$

$$\mathbf{B}'\mathbf{A}'\mathbf{p} \le \mathbf{0}. \tag{3.7}$$

3.2. Simple tree order

Consider the following hypothesis.

$$H_{a_2} : \mathbf{p}^{(\ell)} \leq_S \mathbf{p}^{(g)}, \qquad \text{for } \ell = 1, \ldots, g-1. \tag{3.8}$$

This is a simple tree ordering. Define $k \times k$ matrices $\mathbf{C}_i$'s such that

$$(\mathbf{C}_1', \mathbf{C}_i' \ldots, \mathbf{C}_k')' = \mathbf{C},$$

where $\mathbf{C}$ in defined in simple ordering case. Let

$$\mathbf{A} = [A_1, \ldots, A_{g-1}],$$

where

$$A_i = \underbrace{\mathbf{1} \otimes \cdots \otimes \mathbf{1}}_{i-1} \otimes \begin{bmatrix} \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes \mathbf{C}_1 \\ \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes \mathbf{C}_2 \\ \vdots \\ \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes \mathbf{C}_k \end{bmatrix}$$

and the number of vectors $\mathbf{1}$'s involved in the Kronecker product inside the bracket is $g - i - 1$. We note that the matrix $\mathbf{A}$ and that in simple ordering case are not the same. Then restrictions (3.8) is re-expressed by

$$\mathbf{A}'\mathbf{p} \leq \mathbf{0}$$

and by eliminating redundancy we have

$$\mathbf{B}'\mathbf{A}'\mathbf{p} \leq \mathbf{0}.$$

Finally, we state here an unrestricted alternative hypothesis which is

$$H_{a_0} : \mathbf{p}^{(i)} \neq \mathbf{p}^{(j)}, \qquad \text{for at least one pair of } (i, j) \text{ such } i \neq j. \tag{3.9}$$

Then (3.9) is re-expressed by $\mathbf{A}'\mathbf{p} \neq \mathbf{0}$, and by eliminating redundancy we have $\mathbf{B}'\mathbf{A}'\mathbf{p} \neq \mathbf{0}$. However we are not going to consider the test against this alternative hypothesis.

4. Test Statistic and It's Distribution

Now we are going to compute the likelihood ratio test statistic. Derivation of test statistics rely heavily on Fenchel duality and Khun-Tucker condition. Details of derivation is omitted here but interested reader may refer Jordan (1999) for $g = 2$ case. The extension to the case of $g > 2$ is straightforward.

Suppose $\hat{p}_{i_1 i_2 \ldots i_g} > 0$ for all $i_\ell = 1, 2, \ldots, k$ and $\ell = 1, \ldots, g$. Let $\mathbf{W}^{-1}_{k^g \times k^g} = \operatorname{diag}\{1/\hat{p}_{i_1 i_2 \ldots i_g}\} - \hat{\mathbf{p}}\hat{\mathbf{p}}'$. Then the likelihood ratio test rejects the null for the large value of T,

$$\begin{aligned} T &= -2 \ln \left\{ \frac{\sup_{\mathbf{p} \in H_0} \prod_{i_\ell = 1, \ldots, k, \ell = 1, \ldots, g} p_{i_1 i_2 \ldots i_g}^{n_{i_1 i_2 \ldots i_g}}}{\sup_{\mathbf{p} \in H_a} \prod_{i_\ell = 1, \ldots, k, \ell = 1, \ldots, g} p_{i_1 i_2 \ldots i_g}^{n_{i_1 i_2 \ldots i_g}}} \right\} \\ &= n \min_{\alpha \leq \mathbf{0}} \left\{ \left[\left(\mathbf{B}'\mathbf{A}'\mathbf{W}^{-1}\mathbf{A}\mathbf{B} \right)^{-1} \mathbf{B}'\mathbf{A}'\hat{\mathbf{p}} - \alpha \right]' \left(\mathbf{B}'\mathbf{A}'\mathbf{W}^{-1}\mathbf{A}\mathbf{B} \right) \left[\left(\mathbf{B}'\mathbf{A}'\mathbf{W}^{-1}\mathbf{A}\mathbf{B} \right)^{-1} \mathbf{B}'\mathbf{A}'\hat{\mathbf{p}} - \alpha \right] \right\}, \end{aligned} \tag{4.1}$$

where $\alpha' = (\alpha_1, \ldots, \alpha_{(g-1)(k-1)})$. We note that T can be obtained via quadratic programming. There is substantial literature concerning quadratic programming, for example, Sposito (1975).

To find a critical value for the test, we need to know the asymptotic null distribution of test statistic T. For $g = 2$, El Barmi and Dykstra (1995) and Jordan (1999) showed that the asymptotic null distribution of T is a chi-bar-square distribution. Unfortunately, for $g > 2$, we are unable to derive the exact distribution at this moment. But we are sure that this distribution is going to be a chi-bar-square distribution. That is, for $t > 0$,

$$\begin{aligned}\lim_{n\to\infty} P(T \geq t) &= \sum_{\ell=0}^{(g-1)(k-1)} w(\ell, \mathbf{p}; C) P\left(\chi^2_{(g-1)(k-1)-\ell} \geq t\right) \\ &\leq \frac{1}{2}\left\{P\left(\chi^2_{(g-1)(k-1)} \geq t\right) + P\left(\chi^2_{(g-1)(k-1)-1} \geq t\right)\right\}, \qquad (4.2)\end{aligned}$$

where $w(\ell, \mathbf{p}; C)$ is a level probability and C is a convex cone of dimension $(g-1)(k-1)$. See Robertson *et al.* (1988) for definition and computation of level probability. Specifically, we were unable to find exact expression of these level probabilities $w(\ell, \mathbf{p}; C)$. On the other hand, we may use (4.2) to find a critical value for somewhat conservative test. Note that (4.2) is so-called least favorable distribution.

Feng and Wang (2007) studied likelihood ratio test against simple stochastic ordering among several multinomial distribution. They gave the asymptotic null distribution of test statistic which is a chi-bar-square distribution. We may use this result to find a critical value, but for the large value of g and k the values of level probabilities are intractable. Wang (1996) suggested bootstrapping method. For the small values of g and k we may consider permutation tests.

References

Dardanoni, V. and Forcina, A. (1998). A unified approach to likelihood inference on stochastic orderings in a nonparametric context, *Journal of the American Statistical Association,* **93**, 1112–1123.

Donner, A., Shoukri, M., Klar, N. and Bartfay, E. (2000). Testing the equality of two dependent kappa statistics, *Statistics in Medicine,* **19**, 373–387.

El Barmi, H. and Dykstra, R. (1995). Testing for and against a set of linear inequality constraints in a multinomial setting, *The Canadian journal of Statistics,* **23**, 131–143.

El Barmi, H. and Johnson, M. (2006). A unified approach to testing for and against a set of linear inequality constraints in the product multinomial setting, *Journal of Multivariate Analysis,* **97**, 1894–1912.

El Barmi, H. and Mukerjee, H. (2005). Inferences under a stochastic ordering constraint: The K-sample case, *Journal of the American Statistical Association,* **100**, 252–261.

Feng, Y. and Wang, J. (2007). Likelihood ratio test against simple stochastic ordering among several multinomial populations, *Journal of Statistical Planning and Inference,* **137**, 1362–1374.

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

Gilula, Z. and Haberman, S. J. (1995). Dispersion of categorical variables and penalty functions; derivation, estimation and comparability, *Journal of the American Statistical Association,* **90**, 1447–1452.

Jordan, J. L. (1999). A test of marginal homogeneity versus stochastic ordering in contingency tables, Ph. D. Thesis, The University of Iowa.

McKenzie, D. P., MacKinnon, A. J., Péladeau, N., Onghena, P., Bruce, P. C., Clarke, D. M., Harrigan, S. and McGorry, P. D. (1996). Comparing correlated kappas by resampling: Is one level of agreement significantly different from another?, *Journal of Psychiatric Research,* **30**, 483–492.

Oh, M. (2008). Comparison of two dependent agreements using test of marginal homogeneity, *Communications of the Korean Statistical Society,* **15**, 605–614.

Oh, M. (2009). Inference on measurements of agreement using marginal association, *Journal of the Korean Statistical Society,* **38**, 41–46.

Robertson, T. and Wright, F. T. (1981). Likelihood ratio tests for and against a stochastic ordering between multinomial populations, *The Annals of Statistics,* **9**, 1248–1257.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference,* Wiley, Chichester.

Sposito, V. A. (1975). *Linear and Nonlinear Programming,* Iowa State University Press, Ames.

Wang, Y. (1996). A likelihood ratio test against stochastic ordering in several populations, *Journal of the American Statistical Association,* **91**, 1676–1683.

Received October 2009; Accepted November 2009