Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 1 pp. 263~271

A New Agreement Measure for Interval Multivariate Observations

Yonghwan Um[1)]

Abstract

This article presents a new measure of chance-corrected interobserver agreement among multivariate ratings of many observers. Modifying an approach by Berry and Mielke, a new agreement measure is proposed. The important modificaton is to use the volume of simplex composed of data points as the disagreement masure. The proposed measure accounts agreement for multivariate interval observations among many observers. Hypothetical and real-life data sets are analyzed for illustrative purpose.

Keywords : Chance-corrected interobserver agreement, Many observers multivariate interval observations

1. Introduction

The measure of agreement between two or more observers is one of the statistical concerns in educational and psychological research. Many measures of agreement have been proposed for the case of two independent observers per subject with respect to dichotomous outcome. The most popular measure of this type is Cohen's kappa (1960) which was originally introduced as a chance-corrected measure of agreement between two observers for nominal scales. Cohen adjusted the gross agreement, simply the proportion of times that the two observers agree, by considering the extent of agreement that would occur by chance. Chance agreement is defined as the proportion of times that the two observers would be expected to agree if their ratings were independent of each other. Consider the results of 2 by 2 cross-classification as summarized in Table 1. Cohen's kappa is given by (1) as a function of observed frequencies,

1) Associate Professor, Division of Computer Science, Sungkyul University, Anyang, 430-742, Korea
Email : uyh@sungkyul.edu

$$k = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

where $P_o = P_{11} + P_{22}$ is the observed proportion of observations on which the observers agree, $P_e = P_{1.}P_{.1} + P_{2.}P_{.2}$ is the proportion of observations for which agreement is expected by chance. And in this configuration, $P_0 - P_e$ is the proportion of agreement beyond what is expected by chance, $1 - P_e$ is the maximum possible proportion of agreement beyond what is expected by chance, and the coefficient kappa is the proportion of agreement between two observers after chance agreement is removed.

Table 1. Example of 2 by 2 Cross-classification

observer 1 \ observer 2	1	2	marginals
1	P_{11}	P_{12}	$P_{1.}$
2	P_{21}	P_{22}	$P_{2.}$
marginals	$P_{.1}$	$P_{.2}$	$P_{..} = 1$

Although Cohen's kappa is a most widely used measure of agreement, its use is limited only to the univariate nominal data. Thus several agreement measures to generalize Cohen's kappa to interval data and/or multiple observers have been proposed. The measure proposed by Fleiss(1971) is dependent on the average proportion of observers who agree on the classification of each observation. This measure does not reduce to Cohen's kappa when the number of observers is two. Conger(1980) provided a summary of the problems of extending Cohen's kappa to multiple observers for categorical data. Berry and Mielke(1988) developed an extension of kappa for the case of several observers and one nominal variable, and also for the case of several observers and multivariate interval or ordinal data. Recently, Janson and Olsson(2001) proposed an agreement measure for multivariate interval or nominal data by modifying Berry and Mielke(1988)'s approach.

The purpose of this paper is to propose a new agreement measure for multivariate interval data among several observers. The proposed measure is obtained by modifying the approach by Berry and Mielke(1988) and by using the concept of simplex composed of data points.

2. Berry and Mielke's Agreement Measure.

Berry and Mielke defined their agreement measure as

$$R = 1 - \frac{\delta}{\mu_\delta} \tag{2}$$

where $\delta = 1 - P_0$ represents the observed proportion of disagreement and $\mu_\delta = 1 - P_e$ represents the expected proportion of disagreement. They interpreted kappa as a ratio of measures of disagreement which is measured by the Euclidean distance between the classification of the two observers. Here δ is given by

$$\delta = \left[n\binom{b}{2} \right]^{-1} \sum_{i=1}^{n} \sum_{s<t} \triangle(\mathbf{x}_{si}, \mathbf{x}_{ti}) \tag{3}$$

where b is the number of observers and $\sum_{s<t}$ is sum over all s and t such that $1 \le s < t \le b$ with $\triangle(\mathbf{x}_{si}, \mathbf{x}_{ti}) = \left[\sum_{k=1}^{c} (x_{sik} - x_{tik})^2 \right]^{1/2}$ where $x_{sik}(x_{tik})$ denotes the *k*th element of vector $\mathbf{x}_{si}(\mathbf{x}_{ti})$ with dimension c and i = 1, 2, . . .n(for observations through n). That is, δ is the average(over objects and pairs of observers) of the Euclidean distance between observers' ratings of the same objects. And μ_δ is defined as

$$\mu_\delta = \left[n^2\binom{b}{2} \right]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{s<t} \triangle(\mathbf{x}_{si}, \mathbf{x}_{tj}) \tag{4}$$

with $\triangle(\mathbf{x}_{si}, \mathbf{x}_{tj}) = \left[\sum_{k=1}^{c} (x_{sik} - x_{tjk})^2 \right]^{1/2}$ and signifies the average of the Euclidean distance between one observer's rating of an object and any other observer's rating of any object.

The measure (*R*) satisfies the desired property of chance-corrected measure and is allpicable to interval(and ordinal) data and to multiple observers. Note that, in this representation, Berry and Mielke use a pairwise definition of agreement and take an average of all agreement measures coming from $\binom{b}{2}$ pairs of observers in order to calculate overall agreement measure.

3. New Measure of Agreement

The important modification of Berry and Mielke's approach is that we propose to use volume of simplex as the disagreement measure (rather than Euclidean distance). Consider a bivariate data $\mathbf{x}_{v1}, \ldots, \mathbf{x}_{vn}$, v = 1, 2 , . . . , b, from n objects rated by b observers. For object i, given any three data points $\mathbf{x}_{ri}$, $\mathbf{x}_{si}$

and $\mathbf{x}_{ti}$, we can form the closed triangle with vertices $\mathbf{x}_{ri}$, $\mathbf{x}_{si}$ and $\mathbf{x}_{ti}$. In this way we generate $\binom{b}{3}$ triangles from the b observers.

Observed proportion of disagreement, d_0, is defined in the form of the

$$d_0 = \left[n\binom{b}{3}\right]^{-1} \sum_{i=1}^{n} \sum_{r<s<t} \triangle(\mathbf{x}_{ri}, \mathbf{x}_{si}, \mathbf{x}_{ti}) \tag{5}$$

where $\sum_{r<s<t}$ is sum over all r, s and t such that $1 \le r < s < t \le b$ and

$$\triangle(\mathbf{x}_{ri}, \mathbf{x}_{si}, \mathbf{x}_{ti}) = \frac{1}{2!} \mathrm{abs}\left(\begin{vmatrix} 1 & 1 & 1 \\ x_{ri1} & x_{si1} & x_{ti1} \\ x_{ri2} & x_{si2} & x_{ti2} \end{vmatrix}\right)$$

is the volume of the tringle with vertices $\mathbf{x}_{ri}$, $\mathbf{x}_{si}$ and $\mathbf{x}_{ti}$. Thus, d_0 is interpreted as the average of the volume of triangle composed of observers' ratings of the same objects.

Expected proportion of disagreement, d_e, is average of the volume of triangle composed of one observer's rating of an object and any other observer's rating of any object. This can be expressed in the following form

$$d_e = \left[n^3 \binom{b}{3}\right]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{r<s<t} \triangle(\mathbf{x}_{ri}, \mathbf{x}_{sj}, \mathbf{x}_{tk}) \tag{6}$$

with

$$\triangle(\mathbf{x}_{ri}, \mathbf{x}_{sj}, \mathbf{x}_{tk}) = \frac{1}{2!} \mathrm{abs}\left(\begin{vmatrix} 1 & 1 & 1 \\ x_{ri1} & x_{sj1} & x_{tk1} \\ x_{ri2} & x_{sj2} & x_{tk2} \end{vmatrix}\right)$$

Then the agreement measure denoted by ϕ is defined as

$$\phi = 1 - \frac{d_o}{d_e} \tag{7}$$

Let now, more generally, $\mathbf{x}_{v1}, \ldots, \mathbf{x}_{vn}$, v = 1, 2, . . . , b, be a c-variate data. Then, observed(d_0) and expected(d_e) proportions of disagreement are

$$d_0 = \left[n\binom{b}{w}\right]^{-1} \sum_{i=1}^{n} \sum_{1 \le s_1 < \ldots < s_w \le b} \triangle(\mathbf{x}_{s_1 i}, \mathbf{x}_{s_2 i}, \ldots, \mathbf{x}_{s_w i}) \tag{8}$$

and

$$d_e = \left[n^w \binom{b}{w}\right]^{-1} \sum_{i_1=1}^{n} \cdots \sum_{i_w=1}^{n} \sum_{1 \le s_1 < \ldots < s_w \le b} \triangle(\mathbf{x}_{s_1 i_1}, \mathbf{x}_{s_2 i_2}, \ldots, \mathbf{x}_{s_w i_w}), \tag{9}$$

respectively, where w = c+1 and $\triangle(\mathbf{x}_{s_1 i_1}, \mathbf{x}_{s_2 i_2}, \ldots, \mathbf{x}_{s_w i_w})$ is the volume of the simplex with vertices $\mathbf{x}_{s_1 i_1}, \mathbf{x}_{s_2 i_2}, \ldots, \mathbf{x}_{s_w i_w}$. As in the bivariate case the volume is given by

$$\triangle(\mathbf{x}_{s_1 i_1}, \mathbf{x}_{s_2 i_2}, \ldots, \mathbf{x}_{s_w i_w}) = \frac{1}{c!}\,\mathrm{abs}\left(\begin{vmatrix} 1 & 1 & \cdots & 1 \\ x_{s_1 i_1 1} & x_{s_2 i_2 1} & \cdots & x_{s_w i_w 1} \\ x_{s_1 i_1 2} & x_{s_2 i_2 2} & \cdots & x_{s_w i_w 2} \\ \vdots & \vdots & \cdots & \vdots \\ x_{s_1 i_1 c} & x_{s_2 i_2 c} & \cdots & x_{s_w i_w c} \end{vmatrix}\right)$$

Note that the measure (ϕ) is an average of all $\binom{b}{w}$ agreement measures, each one of which is calculated among w observers. When b = w, ϕ is the omnibus agreement measure among all observers.

When agreement is measured multivariately, one important consideration is the variation of each variable or dimension. Because variables are measured on different scales, they perhaps have unequal variances and contribute differently to observed and expected disagreement. Standardization is then employed to equalize the variances of all variables. All observers' ratings on a variable are considered when standardization is performed. But it should be mentioned that the proposed measure (ϕ) in equation (7) is independent of the units of variables because

$$\triangle(\mathbf{x}_1^{st}, \mathbf{x}_2^{st}, \ldots, \mathbf{x}_w^{st}) = a * \triangle(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_w)$$

where $\mathbf{x}_1^{st}, \mathbf{x}_2^{st}, \ldots, \mathbf{x}_w^{st}$ are standardized vector and the constant factor, $a \neq 0$, appearing both in numerator and denominator of equation (7) is cancelled out.

4. Comparison Between ϕ and R

In order to compare ϕ with R, we used hypothetical data of five objects. Let the bivariate data of (65, 170), (70, 175), (75, 178), (80, 182), and (85, 187), denoted by (wt_i , ht_i), i = 1, 2, . . 5, be hypothetical observations of weight and height rated by observer 1. And let ($wt_i + \triangle$, ht_i) and (wt_i , $ht_i + \triangle$) with $\triangle > 0$ for all i = 1, 2, . . 5, be the data of observer 2 and observer 3, respectively.

We compared ϕ with R by changing the values of $\triangle$ of $m(0 \leq m \leq 5)$ observations. For example, Figure 1(a) shows the agreement measures ϕ and R computed when the data (wt_i+1, ht_i) and (wt_i, ht_i+1) of m observations(first m observations of 5) change to (wt_i+3, ht_i) and (wt_i, ht_i+3), correspondingly, i.e. the disagreements of m observations among the observers increase. For all cases of Figure 1, Berry and Mielke's agreement measure, R, unusually increases when the disagreement increases to m=1 (or m=1 and 2), whereas ϕ steadily decreases as we expect. So ϕ is better agreement measure to use than R when

disagreement is small. At the point of m=3, ϕ and R give similar values of agreement measure. When disagreement is large(m=4 and 5), R reflects the disagreement better than ϕ does. Thus, we can say that ϕ is recommendable to use when there are more observations with small disagreement among observers than the ones with big disagreement.

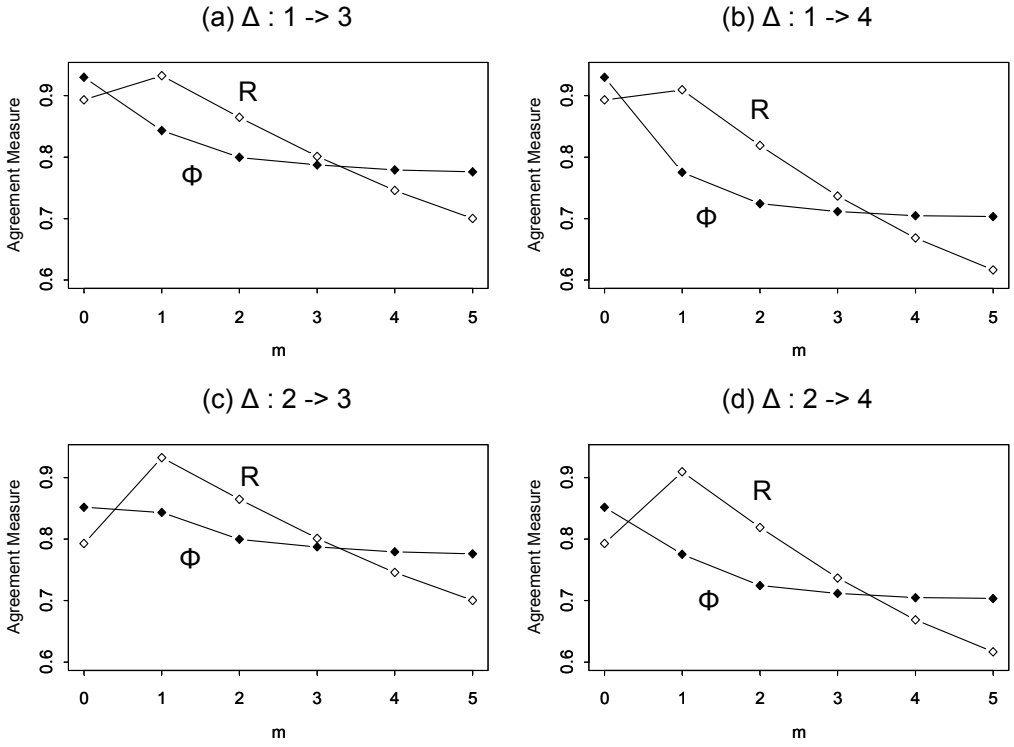

Figure 1. Comparison of Φ with R

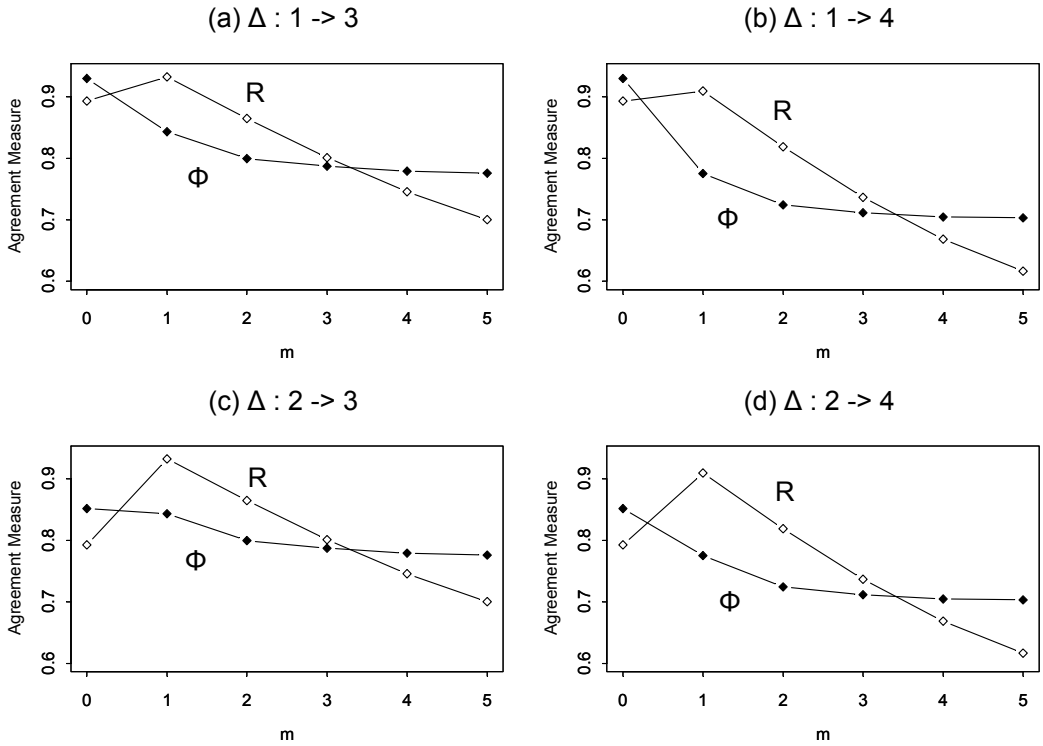

Figure 1. Comparison of Φ with R

5. Example

To illustrate the calculation of agreement measure, we considered another hypothetical bivariate data in Table 2. Three observers rated height and weight of seven men on the basis of photographs. Based on the data in Table 2, observed disagreement, d_0, is 41.143 using equation (5) and expected disagreement, d_e, is 115.960 using equation (6). Inserting the values of d_0 and d_e into equation (7) yields an agreement measure, ϕ, of 0.645.

Table 2. Ratings of weight and height

object	observer 1		observer 2		observer 3	
	weight	height	weight	height	weight	height
1	70	166	76	171	73	170
2	72	160	78	170	78	165
3	85	187	91	174	100	185
4	57	161	64	163	60	162
5	70	172	75	182	80	181
6	66	175	71	179	73	180
7	66	175	70	178	75	180

As an example of a real-life application of the multivariate approach, we used bivariate interval observations of preschool children's behavior. Table 3 represent three observers' ratings of 8 children in terms of their positive attitude and negative attitude. that are observed during the given period of time. Using equation (5), (6) and (7) gives agreement measure of 0.859.

Table 3. Ratings of eight children's behavior

object		1	2	3	4	5	6	7	8
observer 1	positive	26	33	17	37	33	4	39	1
	negative	14	7	23	3	7	36	1	39
observer 2	positive	24	25	16	38	35	3	40	2
	negative	16	15	24	2	5	37	0	38
observer 3	positive	25	35	15	39	36	5	38	0
	negative	15	15	25	1	4	35	2	40

6. Conclusion

We used the notion of simplex of data points to define the new agreement measure. The proposed measure, ϕ, is applcable to many observers' multivariate observations on interval scale. The utilization of simplex is important in that it gives the same value of agreement measure regardless of the units of variables used.

For the hypothetical data, the new agreement measure, ϕ, showed better performance than Berry and Mielke's agreement measure when the observations with small disagreement are dominant among data.

For the example of data in section 5, the multivariate approach makes it possible to assess agreement at an aggregate level, that is, about how well the observers agreed on the mutivariate-rating task as a whole.

In future work, we will address the case with multivariate nominal or ordinal

data which are not covered in this paper and assess the variability of ϕ through bootstrapping.

References

1. Berry, K. J. and Mielke, P. W. Jr. (1988). A Generalization of Cohen's Kappa Agreement Measure to Interval Measurement and Multiple Raters. *Educational and Psychological Measurement,* 48, 921-933
2. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, 20, 37-46.
3. Conger, A. J. (1980). Integration and Generalization of Kappa for Multiple Raters. *Psychological Bulletin*, 88, 322-328
4. Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters, *Psychological Bulletin*, 76, 378-382.
5. Janson, H. and Olsson, U. (2001). A Measure of Agreement for Interval or Nominal Multivariate Observations, *Educational and Psychological Measurement*, 61, 2, 277-289.

[received date : Oct. 2003, accepted date : Feb. 2004]