Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 3, pp. 621~628

A Joint Agreement Measure Between Multiple Raters and One Standard

Yonghwan Um[1)]

Abstract

This article addresses the problem of measuring a joint agreement between multiple raters and a standard set of responses. A new agreement measure based on Um's approach is proposed. The proposed agreement measure is used for multivariate interval responses. Comparison is made between the proposed measure and other corresponding agreement measures using hypothetical data set.

Keywords : multiple raters and a standard set of responses, multivariate interval responses

1. Introduction

The measure of agreement between two or more observers is one of the statistical concerns in educational and psychological research. The most popular measure of this type is Cohen's kappa (1960), where a value of one indicates perfect agreement and zero indicates only chance agreement. This measure has been generalized to multiple raters and high level(ordinal and interval) data by several investigators (Light, 1971; Landis and Koch, 1977; Berry and Mielke, 1988; Janson and Olsson, 2001; Um, 2004). Light(1971) extended Cohen's kappa to multiple raters that is based on the average of all pair-wise kappas. Landis and Koch(1977) considered agreement among several observers in terms of a majority opinion. Berry and Mielke(1988), Janson and Olsson(2001) and Um(2004) proposed agreement measures among many observers applicable to multivariate interval data.

Berry and Mielke(1988) defined their agreement measure by applying a

1) Associate professor, Division of e-business IT, Sungkyul University, Anyang, 430-742, Korea
Email : uyh@sungkyul.edu

multivariate randomized block design with raters as blocks, and extended Cohen's kappa to several raters and one nominal variable and also to several raters and multivariate interval or ordinal data. Janson and Olsson(2001) proposed an agreement measure for multivariate interval or nominal data by modifying Berry and Mielke's(1988) approach. Their modification is to utilize the squared Euclidean distance as disagreement measure rather than Euclidean distance used for Berry and Mielke's(1988) approach. For both of the cases (Berry and Mielke(1988), Janson and Olsson(2001)), they all defined an agreement measure as 1-(observed disagreement/expected disagreement). Um(2004) also proposed a new agreement measure among a set of several observers for multivariate interval data. Um(2004) used a volume of c-dimensional simplex composed of data points as the disagreement measure. Um's(2004) agreement measure for c - variate interval data, ϕ, is expressed as $\phi = 1 - v_o / v_e$, where v_0 is the observed disagreement and v_e is the expected disagreement.

The purpose of this paper is to propose a joint agreement measure between multiple raters and a standard or "correct set" of responses(which may be one of the raters). It is often necessary to measure the degree of agreement between multiple raters and a standard set of responses rather than among many raters simply. Guetzkow(1950) proposed procedures to measure agreement among multiple raters coding items when each item belongs to a known category. Light(1971) provided a test for a joint agreement of multiple raters with a correct set of responses. Hubert(1977) introduced a "target rater" measure of agreement that is identical to the measure proposed by Light(1971). Recently, Berry and Mielke's(1997) proposed a generalized measure of agreement between multiple raters and a set of correct responses at any level of measurement and among multiple responses. In this article, a statistics of agreement based on Um's(2004) measure is presented that measures the agreement of multiple raters with standard set of responses. The proposed agreement measure is used for multivariate interval responses. The use of the proposed measure is exemplified with hypothetical data set and the proposed measure is compared with the measure proposed by Berry and Mielke's(1997) and the corresponding measure that builds on Janson and Olsson's(2001) approach, respectively.

2. Agreement Measure between Multiple Raters and a Standard

Let c be the number of responses/dimensions for each of n objects scored by b raters, and denote the index of the standard set by s. Then the expression for observed disagreement, v_0, in Um's(2004) agreement measure, ϕ, is given by

$$v_0 = \sum_{i=1}^{m} (v_0)_i \tag{1}$$

where $(v_0)_i = \frac{1}{n} \sum_{j=1}^{n} \triangle(i;j)$ is the ith observed disagreement (among c raters and the standard set, s) representing the average, over objects combinations of c +1 raters, of $\triangle(i;j)$'s among raters' observations of the same objects, $m = \binom{b}{c}$ and $\triangle(i;j)$ is the volume of simplex (defined by c data points and a standard set, s) which is calculated as a determinant of matrix formed by c data points and the standard set, s. For example, consider a bivariate data $\mathbf{x}_{p1}, \mathbf{x}_{p2}, .., \mathbf{x}_{pn}$, p = 1, 2 and 3 from n objects rated by 3 raters and a standard set of responses $\mathbf{x}_{s1}, \mathbf{x}_{s2}, .., \mathbf{x}_{sn}$. Then there are three volumes of simplexes to be formed, $\triangle(1;j)$ (defined by s and p=1,2), $\triangle(2;j)$(defined by s and p=1,3) and $\triangle(3;j)$ (defined by s and p=2,3). Here $\triangle(1;j)$ (similarly for $\triangle(2;j)$ and $\triangle(3;j)$) is given by

$$\triangle(1;j) = \frac{1}{2!} \text{ abs} \left(\begin{vmatrix} 1 & 1 & 1 \\ x_{sj1} & x_{1j1} & x_{2j1} \\ x_{sj2} & x_{1j2} & x_{2j2} \end{vmatrix} \right).$$

The expected disagreement, v_e , in Um's(2004) agreement measure(we denote it by UM) is given by

$$v_e = \sum_{i=1}^{m} (v_e)_i \tag{2}$$

where $(v_e)_i = \frac{1}{n^{c+1}} \sum_{j_1=1}^{n} \cdots \sum_{j_{c+1}=1}^{n} \triangle(i;j_1,\cdots,j_{c+1})$ is the ith expected disagreement representing the average, over objects and combinations of c +1 raters, of $\triangle(i;j_1,\cdots,j_{c+1})$'s among raters' observations of any object and $\triangle(i;j_1,\cdots,j_{c+1})$ is the volume of simplex (defined by c data points and a standard set, s). For the same bivariate data above, $\triangle(1;j_1,j_2,j_3)$ (defined by s and p=1,2) is given by

$$\triangle(1;j_1,j_2,j_3) = \frac{1}{2!} \text{ abs} \left(\begin{vmatrix} 1 & 1 & 1 \\ x_{sj_11} & x_{1j_21} & x_{2j_31} \\ x_{sj_12} & x_{1j_22} & x_{2j_32} \end{vmatrix} \right).$$

(Corresponding expressions are also given for $\triangle(2;j_1,j_2,j_3)$ and $\triangle(3;j_1,j_2,j_3)$ similarly).

In order to compare ϕ with corresponding measures that are based on Berry and Mielke's(1997) approach and Janson and Olsson's(2001) approach, respectively, we used the following expressions for the observed disagreement and the expected disagreement in their agreement measures.

$$\text{observed disagreement} = \sum_{i=1}^{b}(t_0)_i, \tag{3}$$

where $(t_0)_i = \frac{1}{n}\sum_{j=1}^{n} D(i;j)$ and

$$\text{expected disagreement} = \sum_{i=1}^{b}(t_e)_i, \tag{4}$$

where $(t_e)_i = \frac{1}{n^2}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} D(i;j_1,j_2)$.

Here $D(i;j)$ is the Euclidean distance $\left(=\sqrt{\sum_{k=1}^{c}(x_{sjk}-x_{ijk})^2}\right)$ for Berry and Mielke's(1997) agreement measure(we denote it by BM) and the squared Euclidean distance $\left(=\sum_{k=1}^{c}(x_{sjk}-x_{ijk})^2\right)$ for Janson and Olsson's(2001) agreement measure (we denote it by JO). Similarly, $D(i;j_1,j_2)=\sqrt{\sum_{k=1}^{c}(x_{sj_1k}-x_{ij_2k})^2}$ for BM and $D(i;j_1,j_2)=\sum_{k=1}^{c}(x_{sj_1k}-x_{ij_2k})^2$ for JO.

3. Numerical Example

In order to illustrate the calculation of agreement measure, we considered a bivariate data in Table 1. Four observers(a standard and 3 observers) rated height and weight of seven men on the basis of photographs. Based on the data in Table 1, observed disagreement, v_0, is 60.29 using equation (1) and expected disagreement, v_e, is 282.88 using equation (2). Inserting the values of v_0 and v_e into $\phi = 1 - v_o/v_e$ yields an agreement measure(we denote it by UM) of 0.787. For the same data set, Berry and Mielke's agreement measure(BM) and Janson

and Olsson's agreement measure(JO) are calculated as 0.631 and 0.881, respectively.

Table 1. Raters' Observations of Weight and Height

object	standard		observer 1		observer 2		observer 3	
	weight	height	weight	height	weight	height	weight	height
1	71	167	70	166	76	171	73	170
2	73	167	72	160	78	170	78	165
3	90	180	85	187	91	174	100	185
4	61	161	57	161	64	163	60	162
5	76	176	70	172	75	182	80	181
6	70	177	66	175	71	179	73	180
7	71	177	66	175	70	178	75	180

4. Comparison among UM, BM and JO

To make a comparison among UM, BM and JO, we used a 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 a standard set of data of weight and height. And let $(wt_i + d1, ht_i)$, $(wt_i, ht_i + d2)$ and $(wt_i + d1, ht_i + d2)$ for all i = 1, 2, ⋯, 5, be the ratings from observer1, observer2 and observer3, respectively. Comparison among UM, BM and JO is made by varying the ratings from observer1, observer2 and observer3. We let $(wt_i + d1, ht_i)$, $(wt_i, ht_i + d2)$ and $(wt_i + d1, ht_i + d2)$ from three observers of the first t (t =0,1,2, ⋯,5) objects move to $(wt_i + d1 + d3, ht_i)$, $(wt_i, ht_i + d2 + d3)$ and $(wt_i + d1 + d3, ht_i + d2 + d3)$ by d3. That is, disagreements of observers on the ratings of t objects increase as the values of d3 of t (t = 0,1, . ,5) objects increase. For instance, Figure 1(a) shows the agreement measures(UM, BM and JO) computed when $(wt_i + 1, ht_i)$, $(wt_i, ht_i + 1)$ and $(wt_i + 1, ht_i + 1)$ of first t objects change to $(wt_i + 2, ht_i)$, $(wt_i, ht_i + 2)$ and $(wt_i + 2, ht_i + 2)$, correspondingly. Figure 1 and 2 show that all agreement measures, UM, BM and JO, decrease as t changes from 0 to 5 and d3 changes from 1 to 2 (i.e. amount of disagreements increases) as we expect. But JO gives big values of agreement measures at all values of t unlike UM and

BM. Even when there are high disagreements among observers (e.g. t =4, 5), JO still gives big values of showing high agreement among observers. (The smallest value of JO is 0.887 when (d1,d2,d3)=(2,2,2) and t = 5). According to Landis and Koch (1977), such a high magnitude of agreement measure is interpreted as the observers' ratings are in 'almost perfect' agreement. But now the values of JO are too big to apply the interpretation to our case. Thus this indicates that JO inflates agreement measure, which is the same result as we observed in Um(2004). On the other hand, UM and BM give similar values of agreement measures over the whole region of t and the magnitudes of UM and BM are reasonable enough to explain agreement among observers. When (d1,d2,d3)=(2,2,2) and t = 5, UM is 0.599 and BM is 0.605, which can be interpreted as 'moderate agreement' in Landis and Koch (1977). The similarity between UM and BM, as mentioned in Um(2004), results from that $\triangle(i;j)$ in UM and $D(i;j)$ in BM belong to same metric space. The difference of UM and BM from the early study(Um(2004) is that UM and BM behaves almost the same over the whole region from t = 0 to t = 5. In early study(Um(2004)), UM perfomed better than BM at small size of disagreements(t = 0, 1) and BM did better than UM at large size of disagreements(t = 4, 5). As a result, we state that UM and BM perform similarly and better than JO does when we measure a joint agreement between multiple raters and a standard.

5. Conclusion

A joint agreement measure between multiple raters and a standard set of responses for multivariate interval data is presented (The standard set of responses may be the data from one of the raters). It builds on Um's(2004) approach where the volume of simplex defined by data points is used as disagreement measure. The findings from the comparison study using hypothetical data set are similar to the ones in Um(2004) and as follow: over the whole region of t, (1) JO inflates agreement measure (2) the proposed measure UM performs similarly as BM does and better than JO does.

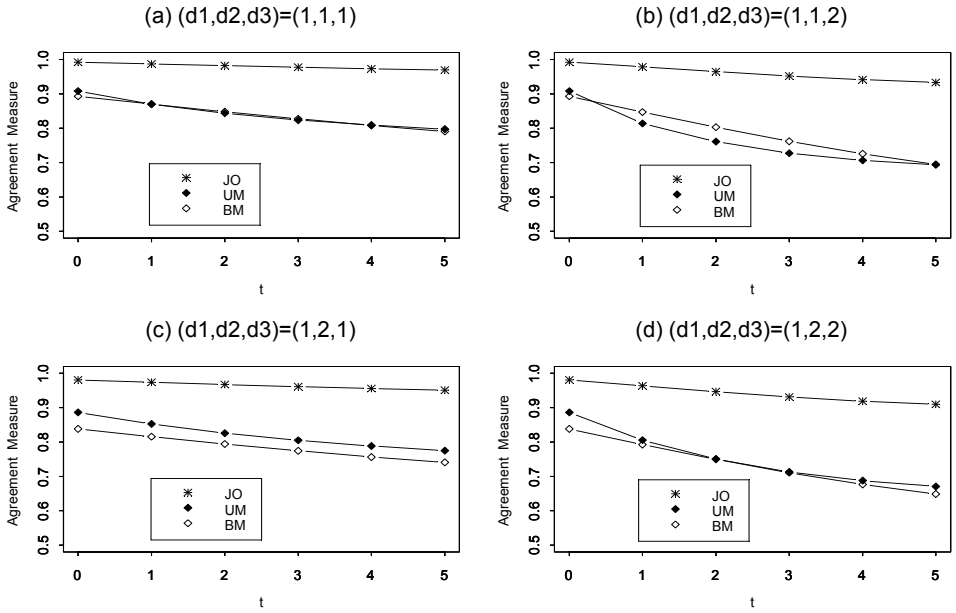

Figure 1. Comparison among UM, BM and JO (d1:1->2 or 3)

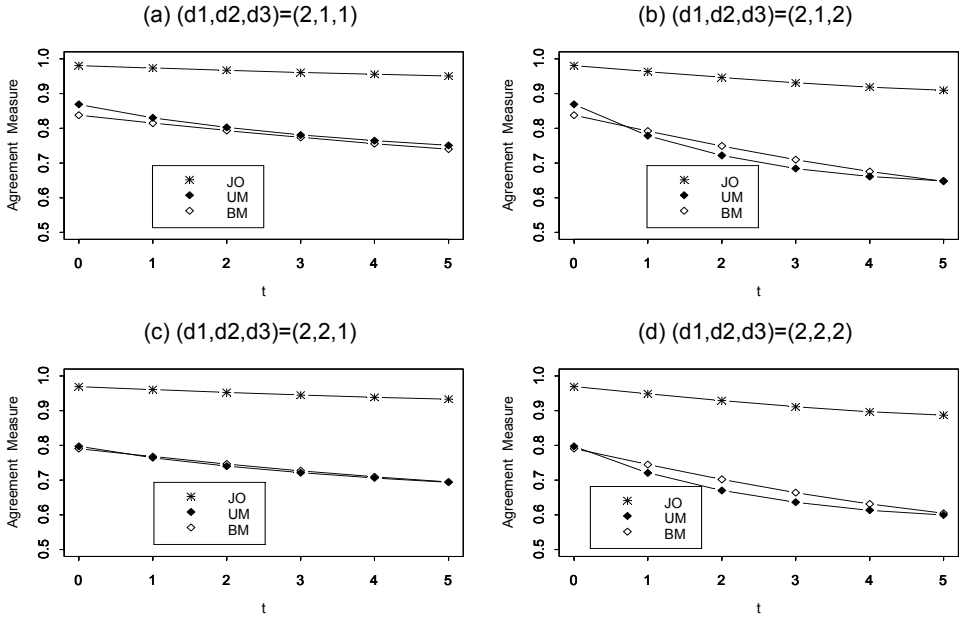

Figure 2. Comparison among UM, BM and JO (d1:2->3 or 4)

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. Berry, K. J., and Mielke, P. W. Jr. (1997). Measuring the Joint Agreement between Multiple Raters and a Standard. *Educational and Psychological Measurement*, *57*, 3, 527-530.
3. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, *20*, 37-46.
4. Janson, H., and Olsson, U. (2001). A Measure of Agreement for Interval or Nominal Multivariate Observations, *Educational and Psychological Measurement*, *61*, 2, 277-289.
5. Guetzkow, H (1950). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *6*, 47-58
6. Hubert, L. (1977). Kappa revisited. *Psychological Bulletin, 84,* 289-297.
7. Light, R. J. (1971). Measures of Response Agreement for Qualitative Data: Some Generalizations and Alternatives. *Psychological Bulletin, 76,* 365-377.
8. Landis, J. R., and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data, *Biometrics, 33,* 159-174.
9. Um, Y. (2004). A New Agreement Measure for Interval Multivariate Observations, *Journal of Korean Data & Information Science Society, 15*, 1, 263-271.

[received date : Apr. 2005, accepted date : May. 2005]