Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 2, pp. 529 ~ 536

On Assessing Inter-observer Agreement Independent of Variables' Measuring Units

Yonghwan Um[1)]

Abstract

Investigators use either Euclidean distance or volume of a simplex defined composed of data points as agreement index to measure chance-corrected agreement among observers for multivariate interval data. The agreement coefficient proposed by Um(2004) is based on a volume of a simplex and does not depend on the variables' measuring units. We consider a comparison of Um(2004)'s agreement coefficient with others based on two unit-free distance measures, Pearson distance and Mahalanobis distance. Comparison among them is made using hypothetical data set.

Keywords : Agreement coefficient, Multivariate interval data, Unit-free distance measure

1. Introduction

The degree to which a group of people share the same opinions (agreement coefficient) is one of the statistical concerns in various observational studies. Especially, many researchers have studied an agreement coefficient in the case where a set of several observers rate a sample of objects multivariatly on several variables(or dimensions). Among them, the studies by Berry and Mielke (1988), Janson and Olsson(2001), and Um(2004) are the most recent ones. The agreement coefficients proposed by them are extensions of Cohen's Kappa(1960) to multiple observers and applicable to multivariate interval data. Berry and Mielke (1988) defined their agreement coefficient on the basis of Euclidean distance between two observations as agreement index whereas Janson and Olsson (2001) utilized the squared Euclidean distance rather than Euclidean distance. The agreement

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

coefficients proposed by Um (2004) are based on the volume of c-dimensional simplex composed of data points whose value is given by the determinant of a matrix. Berry and Mielke (1988), Janson and Olsson (2001) and Um's(2004) all defined their agreement coefficients as 1- (observed disagreement / expected disagreement). All of these agreement coefficients have a property of chance-correctedness which is desirable as an agreement coefficient. Chance-corrected coefficients reflect the amount of agreement in excess of what would be expected by chance (See Brennan and Prediger(1981), Cicchetti, Showalter, and Tyrer(1985), and Conger(1985)). However, as pointed out by authors, the problem with Berry and Mielke's (1988) agreement coefficient and Janson and Olsson's (2001) agreement coefficient is that the contributions made by variables may be different in any one analysis. When variables are measured on different scales, they will contribute differently to the agreement index and this will make agreement coefficient depend on the variables' measuring units used. For example, when observers rate the weight and height of objects on the basis of photographs, observers' agreement on the ratings of weight(measured in kg) and height(measured in cm) is different from that obtained using different units of weight(measured in pound) and height(measured in inch). But Um's(2004) agreement coefficient is independent of variables' measuring units and remains the same regardless of any unit change, which will be proved in section 2.

The purpose of this article is to compare Um's(2004) agreement coefficient and other agreement coefficients defined on the basis of Pearson distance and Mahalanobis distance. Pearson distance and Mahalanobis distance are unit-free distance measures not depending variables' measuring unit unlike the Euclidean distance.

2. Agreement Coefficients Independent of Variables' Measuring Units

(1) Um's Agreement Coefficient's Independence of Measuring Units

Consider a c-variate data $\mathbf{x}_{p1}, \ldots, \mathbf{x}_{pn}$, p = 1, 2, . . . , b, from n objects rated by b observers. The agreement coefficient by Um(2004) is $U = 1 - v_o / v_e$ where v_o, observed proportion of disagreement, is given by

$$v_o = \left[n \binom{b}{c+1} \right]^{-1} \sum_{i=1}^{n} \sum_{1 \le s_1 < \cdots < s_{c+1} \le b} \triangle(\mathbf{x}_{s_1 i}, \mathbf{x}_{s_2 i}, \ldots, \mathbf{x}_{s_{c+1} i}) \qquad (1)$$

and v_e, expected proportion of disagreement, is given by

$$v_e = \left[n^{c+1} \binom{b}{c+1} \right]^{-1} \sum_{i_1=1}^{n} \cdots \sum_{i_p=1}^{n} \sum_{1 \le s_1 < \cdots < s_{c+1} \le b} \triangle(\mathbf{x}_{s_1 i_1}, \mathbf{x}_{s_2 i_2}, \ldots, \mathbf{x}_{s_{c+1} i_{c+1}}) \quad (2)$$

Here, $\triangle(\mathbf{x}_{s_1 i}, \mathbf{x}_{s_2 i}, \ldots, \mathbf{x}_{s_{c+1} i})$ is the volume of the simplex with vertices $\mathbf{x}_{s_1 i}, \mathbf{x}_{s_2 i}, \ldots, \mathbf{x}_{s_{c+1} i}$, which is calculated as a determinant of $(c+1) \times (c+1)$ matrix,

$$\triangle(\mathbf{x}_{s_1 i}, \mathbf{x}_{s_2 i}, \ldots, \mathbf{x}_{s_{c+1} i}) = \frac{1}{c!} \operatorname{abs} \left(\begin{vmatrix} 1 & 1 & \cdots & 1 \\ x_{s_1 i 1} & x_{s_2 i 1} & \cdots & x_{s_{c+1} i 1} \\ x_{s_1 i 2} & x_{s_2 i 2} & \cdots & x_{s_{c+1} i 2} \\ \vdots & \vdots & \cdots & \vdots \\ x_{s_1 i c} & x_{s_2 i c} & \cdots & x_{s_{c+1} i c} \end{vmatrix} \right).$$

Note that $\triangle(\mathbf{x}_{s_1 i}, \mathbf{x}_{s_2 i}, \ldots, \mathbf{x}_{s_{c+1} i})$ is affine invariant under the nonsingular linear transformation of the form $D(x) = Ax+g$ (A is a c by c nonsingular matrix and is g a c by 1 vector) because

$$\triangle(D(\mathbf{x}_{s_1 i}), D(\mathbf{x}_{s_2 i}), \ldots, D(\mathbf{x}_{s_{c+1} i})) = \operatorname{abs}(|A|) \triangle(\mathbf{x}_{s_1 i}, \mathbf{x}_{s_2 i}, \ldots, \mathbf{x}_{s_{c+1} i}).$$

Similarly, we can say that $\triangle(\mathbf{x}_{s_1 i_1}, \mathbf{x}_{s_2 i_2}, \ldots, \mathbf{x}_{s_{c+1} i_{c+1}})$ is also affine invariant. As a result, the utilization of simplex makes U invariant with respect to rotation and reflection and scale transformation. Hence the agreement coefficient, U, dose not depend on variables' measuring units.

(2) Agreement Coefficients Based on Unit-free Distance Measures

We define agreement coefficients corresponding to Pearson distance and Mahalanobis distance that are independent of variables' measuring units. Pearson distance and Mahalanobis distance between two observations, $\mathbf{x}_{si}$ and $\mathbf{x}_{tj}$ are defined as

$$\text{Pearson distance} = \left[\sum_{k=1}^{c} \frac{(x_{sik} - x_{tjk})^2}{V_k^2} \right]^{\frac{1}{2}}, \quad (3)$$

where V_k^2 is the variance of the k-th variable and

$$\text{Mahalanobis distance} = \left[(\mathbf{x}_{si} - \mathbf{x}_{tj})^t \, \mathbf{S}^{-1} (\mathbf{x}_{si} - \mathbf{x}_{tj}) \right]^{\frac{1}{2}}, \quad (4)$$

where $\mathbf{S}$ is the variance-covariance matrix of the sample, respectively. For

Pearson distance measure, standardization is performed to equalize the unequal variances of dimensions that might be due to different scales of measurements. Mahalanobis distance accounts for the correlation as well as the standardization of variables.

Now we express agreement coefficient among a set of observers as agreement coefficient = 1 − observed disagreement / expected disagreement
where

$$\text{observed disagreement} = \left[n\binom{b}{2}\right]^{-1} \sum_{i=1}^{n} \sum_{s<t} \triangle(\mathbf{x}_{si}, \mathbf{x}_{ti}), \tag{5}$$

and

$$\text{expected disagreement} = \left[n^2\binom{b}{2}\right]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{s<t} \triangle(\mathbf{x}_{si}, \mathbf{x}_{tj}) \tag{6}$$

and $\sum_{s<t}$ is sum over all s and t such that $1 \le s < t \le b$. Here $\triangle(\mathbf{x}_{si}, \mathbf{x}_{ti})$ (similarly $\triangle(\mathbf{x}_{si}, \mathbf{x}_{tj})$) is either Pearson distance or Mahalanobis distance between $\mathbf{x}_{si}$ and $\mathbf{x}_{tj}$. We denote agreement coefficient by P (for Pearson distance) and by M (for Mahalanobis distance), respectively.

3. Numerical Example

In order to illustrate the calculation of agreement coefficient, we considered a hypothetical bivariate interval data in Table 1. Three observers rated height and weight of five men on the basis of photographs. Based on the data in Table 1, we first have

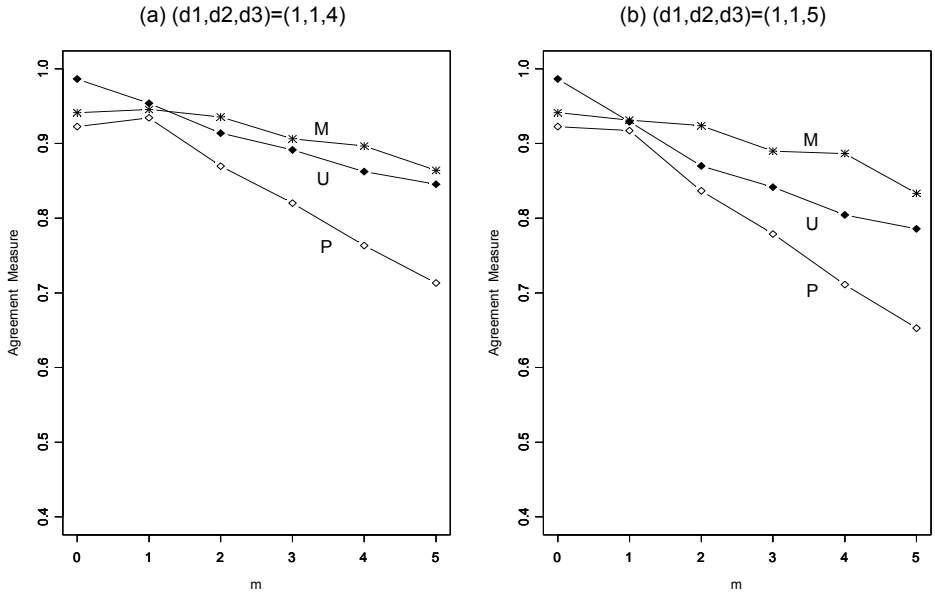
v_0 = 58.6 and v_e = 115.89 using equation (1) and (2) respectively and hence U = 0.494 (with n=5, b=3, and c=2). For the same data set, we have P = 0.481 (with observed disagreement = 0.879 and expected disagreement = 1.694) and M = 0.420 (with observed disagreement = 1.015 and expected disagreement = 1.749).

<Table 1> Observers' Ratings of Weight and Height

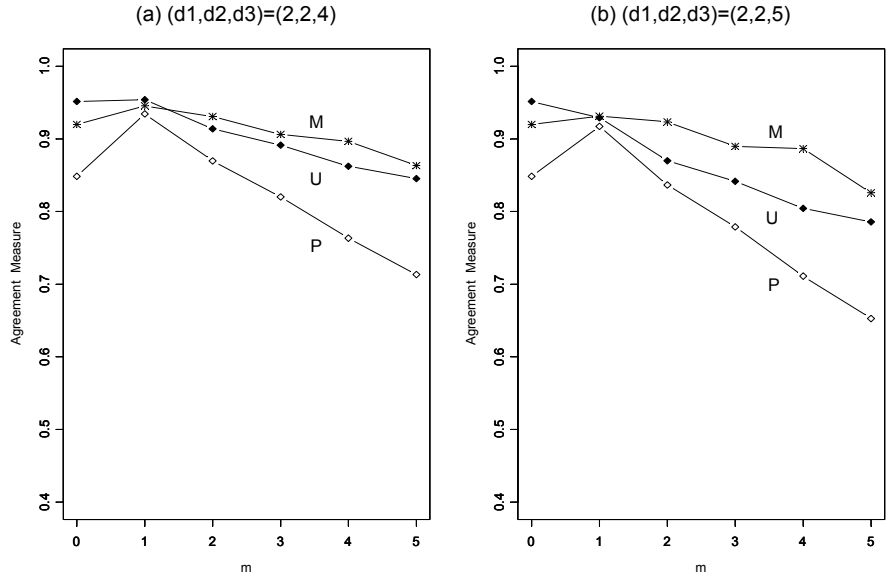
object	observer 1		observer 2		observer 3	
	weight	height	weight	height	weight	height
1	71	166	76	171	74	171
2	73	160	80	170	80	165
3	86	187	93	174	101	185
4	59	161	66	163	62	162
5	71	172	77	182	83	181

4. Comparison among U, P, and M

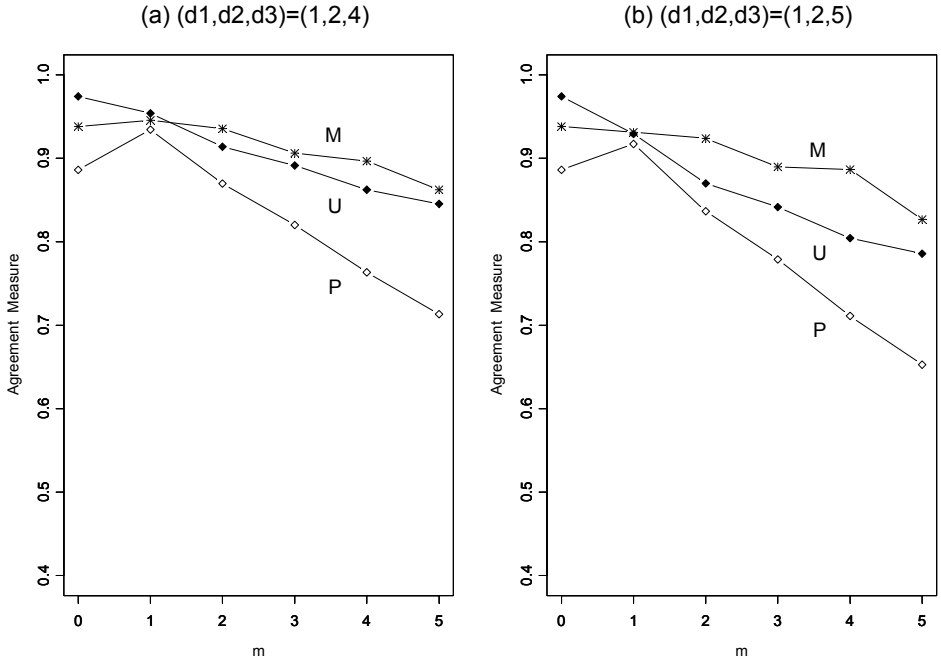
We use hypothetical data of five objects to make a comparison among U, P, and M. Let the bivariate data of (59, 161), (73, 160), (86, 187), (71, 166), and (71, 172), denoted by (wt_i, ht_i), i = 1, 2, . . 5, be hypothetical observations of weight and height rated by observer 1. And let $(wt_i + d1, ht_i)$ and $(wt_i, ht_i + d2)$ for all i = 1, 2, ⋯, 5, be the ratings from observer 2 and observer 3, respectively. Comparison among U, P, and M is made by varying the ratings from observer 2 and observer 3 (while fixing the ratings from observer 1). In order to observe the behaviors of U, P, and M, we let $(wt_i + d1, ht_i)$ and $(wt_i, ht_i + d2)$ from two observers(observer 2 and observer 3) of the first m (m=0, 1, ⋯, 5) objects move to $(wt_i + d1 + d3, ht_i)$, and $(wt_i, ht_i + d2 + d3)$ by d3. That is, observers' disagreements on the ratings of m objects increase as d3 of m (m = 0,1, . ,5) objects increases. Figure 1, 2 and 3 show agreement coefficients(U, P, and M) of three observers with different combinations of (d1, d2, d3). We see that U, P, and M decrease overall as m changes (disagreements increase), and that P shows more rapid change in amount of agreement than others. Note that even when there is high disagreement among observers (e.g. m = 4, 5), M still gives a high value of M (the smallest M is 0.827 when d1=d2=2, m=5) with only a small range(around 0.1) of variation. Such a high magnitude of M, according to Landis and Koch(1977), is interpreted as the observers' ratings are in 'almost perfect' agrement. Thus it appears that M does not behave well enough to detect the high disagreement among observers. But U and P decrease with bigger variation than M as disagreements increase. Especially, P varies considerably as m changes from 0 to 5(e.g. P=0.917 and 0.653 when m=1 and m=5, respectively in Figure 3(b)). This indicates that U and P reflect the disagreement among observers better than M. We also see that P performs better than U as disagreement increases (when $m \geq 2$). However, note that P unusually increases when disagreement increases from m=0 to m=1 (although this similar phenomena occurs also for U, it is much milder than for M). Thus we state that U performs better than P and hence recommendable to use when the disagreements on the observers' ratings are relatively low(m=1). But when the amount of disagreement is moderate or large, P is more useful than others.



<Figure 1> Comparison among U, P, and M (d1=d2=1)



<Figure 2> Comparison among U, P, and M (d1=d2=2)



<Figure 3> Comparison among U, P, and M (d1=1 and d2=2)

5. Conclusion

Comparison among the agreement coefficients(U, P, and M) not depending on variables' measuring scales is made. Their independence of variables' units comes from the use of volume of simplex defined by data points, Pearson distance and Mahalanobis distance as agreement index. Especially, U is affine invariant with respect to rotation, reflection and scale transformation. The comparison study using hypothetical data set shows that U and P perform better than M over the whole region of m. The better performance of U and P than M is in a sense that U and P are able to respond to the change of disagreement appropriately. At the small size of disagreement, U performs better than P. Thus we recommend to use U when there are more observations with small disagreement among observers than the ones with moderate or large disagreement.

For the other case, P is recommendable to use.

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. Brennan, R. L. and Prediger, D. L. (1981). Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-699.
3. Cicchetti, D. V., Showalter, D., and Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrrater reliability: a Monte Carlo investigation. Apllied *Psychological Measurement, 9*. 31-36.
4. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, *20*, 37-46.
5. Conger, A. J. (1985). Kappa reliabilities for continuous behaviors and events. *Educational and Psychological Measurement, 45*, 861-868.
6. Janson, H., and Olsson, U. (2001). A Measure of Agreement for Interval or Nominal Multivariate Observations, *Educational and Psychological Measurement*, *61*, 2, 277-289.
7. Landis, J. R., & Koch, G. G., (1977). The measurement of observer agreement for categorical data, *Biometrics, 33*, 159-174.
8. Um, Y. (2004). A New Agreement Measure for Interval Multivariate Observations, *Journal of Korean Data & Information Science Society, 15*, 263-271.

[received date : Mar. 2006, accepted date : Apr. 2006]