Shanghai Archives of Psychiatry, 2015, Vol. 27, No. 1

•Biostatistics in psychiatry (25)•

Kappa coefficient: a popular measure of rater agreement

Wan TANG[1]*, Jun HU[2], Hui ZHANG[3], Pan WU[4], Hua HE[1,5]

Summary: In mental health and psychosocial studies it is often necessary to report on the between-rater agreement of measures used in the study. This paper discusses the concept of agreement, highlighting its fundamental difference from correlation. Several examples demonstrate how to compute the kappa coefficient – a popular statistic for measuring agreement – both by hand and by using statistical software packages such as SAS and SPSS. Real study data are used to illustrate how to use and interpret this coefficient in clinical research and practice. The article concludes with a discussion of the limitations of the coefficient.

Keywords: interrater agreement; kappa coefficient; weighted kappa; correlation

[*Shanghai Arch Psychiatry*. 2015; **27**(1): 62-67. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.215010]

1. Introduction

For most physical illnesses such as high blood pressure and tuberculosis, definitive diagnoses can be made using medical devices such as a sphygmomanometer for blood pressure or an X-ray for tuberculosis. However, there are no error-free gold standard physical indicators of mental disorders, so the diagnosis and severity of mental disorders typically depends on the use of instruments (questionnaires) that attempt to measure latent multi-faceted constructs. For example, psychiatric diagnoses are often based on criteria specified in the Fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV)[1], published by the American Psychiatric Association. But different clinicians may have different opinions about the presence or absence of the specific symptoms required to determine the presence of a diagnosis, so there is typically no perfect agreement between evaluators. In this situation, statistical methods are needed to address variability in clinicians' ratings.

Cohen's kappa is a widely used index for assessing agreement between raters.[2] Although similar in appearance, agreement is a fundamentally different concept from correlation. To illustrate, consider an instrument with six items and suppose that two raters' ratings of the six items on a single subject are (3,5), (4,6), (5,7), (6,8), (7,9) and (8,10). Although the scores of the two raters are quite different, the Pearson correlation coefficient for the two scores is 1, indicating perfect correlation. The paradox occurs because there is a bias in the scoring that results in a consistent difference of 2 points in the scores of the two raters for all 6 items in the instrument. Thus, although perfectly correlated (precision), there is quite poor agreement between the two raters. The kappa index, the most popular measure of raters' agreement, resolves this problem by assessing both the bias and the precision between raters' ratings.

In addition to its applications to psychiatric diagnosis, the concept of agreement is also widely applied to assess the utility of diagnostic and screening tests. Diagnostic tests provide information about a patient's condition that clinicians' often use when making decisions about the management of patients. Early detection of disease or of important changes in the clinical status of patients often leads to less suffering and quicker recovery, but false negative and false positive screening results can result in delayed treatment or in inappropriate treatment. Thus when a new diagnostic or screening test is developed, it is critical to assess its accuracy by comparing test results with those from a gold or reference standard. When assessing such tests, it is incorrect to measure the correlation of the results of the test and the gold standard, the correct procedure is to assess the agreement of the test results with the gold standard.

[1] Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, United States
[2] College of Basic Science and Information Engineering, Yunnan Agricultural University, Kunming, Yunnan Province, China
[3] Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, United States
[4] Value Institute, Christiana Care Health System, Newark, DE, United States
[5] Center of Excellence for Suicide Prevention, Canandaigua VA Medical Center Canandaigua, NY, United States
*correspondence: wan_tang@urmc.rochester.edu

A full-text Chinese translation of this article will be available at www.shanghaiarchivesofpsychiatry.org/cn on March 25, 2015.

2. Problems

Consider an instrument with a binary outcome, with '1' representing the presence of depression and '0' representing the absence of depression. Suppose two independent raters apply the instrument to a random sample of *n* subjects. Let and denote the ratings on the *n* subjects by the two raters for *i*=1,2,...,*n*. We are interested in the degree of agreement between the two raters. Since the ratings are on the same scale of two levels for both raters, the data can be summarized in a 2×2 contingency table.

To illustrate, Table 1 shows the results of a study assessing the prevalence of depression among 200 patients treated in a primary care setting using two methods to determine the presence of depression;[3] one based on information provided by the individual (i.e., proband) and the other based on information provided by another informant (e.g., the subject's family member or close friend) about the proband. Intuitively, we may think that the proportion of cases in which the two ratings are the same (in this example, 34.5% [(19+50)/200]) would be a reasonable measure of agreement. But the problem with this proportion is that it is almost always positive, even when the rating by the two methods is completely random and independent of each other. So the proportion of overall agreement does not indicate whether or not two raters or two methods of rating are in agreement.

Table 1. Diagnosis of depression among 200 primary care patients based on information provided by the proband and by other informants about the proband

Proband	Informant		total
	not depressed	depressed	
not depressed	66	19	85
depressed	50	65	115
total	116	84	200

For example, suppose that two raters with no training or experience about depression randomly decide whether or not each of the 200 patients has depression. Assume that one rater makes a positive diagnosis (i.e., considers depression present) 80% of the time and the other gives a positive diagnosis 90% of the time. Based on the assumption that their diagnoses are made independently from each other, Table 2 represents the joint distribution of their ratings. The proportion that the two raters give the same diagnosis is 74% (i.e., 0.72+0.02), suggesting that the two raters are doing a good job of diagnosing the presence of depression. But this level of agreement is purely by chance, it does *not* reflect the actual degree of agreement between the two raters. This hypothetical example shows that the proportion of cases in which two raters give the same ratings on an instrument is inflated by the agreement by chance. This chance agreement must be removed in order to provide a valid measure of agreement. Cohen's kappa coefficient is used to assess the level of agreement beyond chance agreement.

Table 2. Hypothetical example of proportional distribution of diagnoses by two raters that make diagnoses independently from each other

Rater 1 result	Rater 2 result		total
	positive	negative	
positive	0.72	0.08	0.80
negative	0.18	0.02	0.20
total	0.90	0.10	1.00

3. Kappa for 2×2 tables

Consider a hypothetical example of two raters giving ratings for *n* subjects on a binary scale, with '1' representing a positive result (e.g., the presence of a diagnosis) and '0' representing a negative result (e.g., the absence of a diagnosis). The results could be reported in a 2x2 contingency table as shown in Table 3. By convention, the results of the first rater are traditionally shown in the rows (x values) and the results of the second rater are shown in the columns (y values). Thus, n_{ij} in the table denotes the number of subjects who receive the rating of *i* from the first rater and the rating *j* from the second rater. Let Pr(A) denote the probability of event A; then p_{ij}=Pr(x=i,y=j) represent the proportion of all cases that receive the rating of *i* from the first rater and the rating *j* from the second rater, p_{i+}=Pr(x=i) represents the marginal distribution of the first rater's ratings, and p_{+j}=Pr(y=j) represents the marginal distribution of the second rater's ratings.

Table 3. A typical 2×2 contingency table to assess agreement of two raters

First rater (x)	Second rater (y)		total
	1 (positive)	0 (negative)	
1 (positive)	n_{11}	n_{10}	n_{1+}
0 (negative)	n_{01}	n_{00}	n_{0+}
total	n_{+1}	n_{+0}	*n*

If the two raters give their ratings independently according to their marginal distributions, the probability that a subject is rated 0 (negative) by chance by both raters is the product of the marginal probabilities p_{0+} and p_{+0}. Likewise, the probability of a subject being rated 1 (positive) by chance by both raters is the product of the marginal probabilities p_{1+} and p_{+1}. The sum of these two probabilities ($p_{1+}*p_{+1} + p_{0+}*p_{+0}$) is the agreement by chance, that is, the source of inflation discussed earlier.

After excluding this source of inflation from the total proportion of cases in which the two raters give identical ratings ($p_{11} + p_{00}$), we arrive at the agreement corrected for chance agreement, ($p_{11}+p_{00} – (p_{1+}*p_{+1} + p_{0+}*p_{+0})$). In 1960 Cohen[1] recommended normalizing this chance-adjusted agreement as the Kappa coefficient (*K*):

$$k = \frac{p_{11} + p_{00} - (p_{1+}p_{+1} + p_{0+}p_{+0})}{1 - (p_{1+}p_{+1} + p_{0+}p_{+0})}. \quad (1)$$

This normalization process produces kappa coefficients that vary between −1 and 1, depending on the degree of agreement or disagreement beyond chance. If the two raters completely agree with each other, then $p_{11}+p_{00}=1$ and $K=1$. Conversely, if the kappa coefficient is 1, then the two raters agree completely. On the other hand, if the raters rate the subjects in a completely random fashion, then the agreement is completely due to chance, so $p_{11}=p_{1+}*p_{+1}$ and $p_{00}=p_{0+}*p_{+0}$ do $(p_{11}+p_{00} – (p_{1+}*p_{+1} + p_{0+}*p_{+0}))=0$ and the kappa coefficient is also 0. In general, when rater agreement exceeds chance agreement the kappa coefficient is positive, and when raters disagree more than they agree the kappa coefficient is negative. The magnitude of kappa indicates the degree of agreement or disagreement.

The kappa coefficient can be estimated by substituting sample proportions for the probabilities shown in equation (1). When the number of ratings given by each rater (i.e., the sample size) is large, the kappa coefficient approximately follows a normal distribution. This asymptotic distribution can be estimated using delta methods based on the asymptotic distributions of the various sample proportions.[4] Based on the asymptotic distribution, calculations of confidence intervals and hypothesis tests can be performed. For a sample with 100 or more ratings, this generally provides a good approximation. However, it may not work well for small sample sizes, in which case exact methods may be applied to provide more accurate inference.[4]

Example 1. *Assessing the agreement between the diagnosis of depression based on information provided by the proband compared to the diagnosis based on information provided by other informants (Table 1), the Kappa coefficient is computed as follows:*

$$k = \frac{\frac{66}{200} + \frac{65}{200} - (\frac{116}{200}\frac{85}{200} + \frac{84}{200}\frac{115}{200})}{1 - (\frac{116}{200}\frac{85}{200} + \frac{84}{200}\frac{115}{200})} = 0.3262.$$

The asymptotic standard error of kappa is estimated as 0.063. This gives a 95% confidence interval of κ, (0.2026, 0.4497). The positive kappa indicates some degree of agreement about the diagnosis of depression between diagnoses based on information provided by the proband versus diagnoses based on information provided by other informants. However, the level of agreement, though statistically significant, is relatively weak.

In most applications, there is usually more interest in the magnitude of kappa than in the statistical significance of kappa. When the sample is relatively large (as in this example), a low kappa which represents relatively weak agreement can, nevertheless, be statistically significant (that is, significantly greater than 0). The degree of beyond-chance agreement has been classified in different ways by different authors who arbitrarily assigned each category to specific cutoff levels of Kappa. For example, Landis and Koch[5] proposed that a kappa in the range of 0.21–0.40 be considered 'fair' agreement, kappa=0.41–0.60 be considered 'moderate' agreement, kappa=0.61–0.80 be considered 'substantial' agreement, and kappa >0.81 be considered 'almost perfect' agreement.

4. Kappa for categorical variables with multiple levels

The kappa coefficient for a binary rating scale can be generalized to cases in which there are more than two levels in the rating scale. Suppose there are *k* nominal categories in the rating scale. For simplicity and without loss of generality, denote the rating levels by 1,2,...,*k*. The ratings from the two raters can be summarized in a $k \times k$ contingency table, as shown in Table 4. In the table, n_{ij}, p_{ij}, p_{i+}, and p_{+j} have the same interpretations as in the 2x2 contingency table (above) but the range of the scale is extended to $i,j=1,\ldots,k$. As in the binary example, we first compute the agreement by chance, (the sum of the products of the *k* marginal probabilities, $\sum p_{i+}*p_{+i}$ for $i=1,\ldots,k$), and subtract this chance agreement from the total observed agreement (the sum of the diagonal probabilities, $\sum p_{ii}$ for $i=1,\ldots,k$) before estimating the normalized agreement beyond chance:

$$k = \frac{\sum_{i=1}^{k} p_{ii} - \sum_{i=1}^{k} p_{i+}p_{+i}}{1 - \sum_{i=1}^{k} p_{i+}p_{+i}}. \quad (2)$$

Table 4. Model KxK contingency table to assess agreement about k categories by two different raters

x	*y*				total
	1	2	...	k	
1	n_{11}	n_{12}	...	n_{1k}	n_{1+}
2	n_{21}	n_{22}	...	n_{2k}	n_{2+}
...	...	...	...	...	...
k	n_{k1}	n_{k2}	...	n_{kk}	n_{k+}
total	n_{+1}	n_{+2}	...	n_{+k}	*n*

As in the case of binary scales, the kappa coefficient varies between −1 and 1, depending on the extent of agreement or disagreement. If the two raters completely agree with each other ($\sum p_{ii}=1$, for $i=1,\ldots,k$), then the kappa coefficient is equal to 1. If the raters rate the subjects at random, then the total agreement is equal chance agreement ($\sum p_{ii} = \sum p_{i+}*p_{+i}$, for $i=1,\ldots,k$) so the

kappa coefficient is 0. In general, the kappa coefficient is positive if there is agreement or negative if there is disagreement, with the magnitude of kappa indicating the degree of such agreement or disagreement between the raters. The kappa index in equation (2) is estimated by replacing the probabilities with their corresponding sample proportions. As in the case of binary scales, we can use asymptotic theory and exact methods to assess confidence intervals and make inferences.

5. Kappa for ordinal or ranked variables

The definition of the kappa coefficient in equation (2) assumes that the rating categories are treated as independent categories. If, however, the rated categories are ordered or ranked (for example, a Likert scale with categories such as '*strongly disagree', 'disagree', 'neutral', 'agree',* and '*strongly agree*'), then a weighted kappa coefficient is computed that takes into consideration the different levels of disagreement between categories. For example, if one rater 'strongly disagrees' and another 'strongly agrees' this must be considered a greater level of disagreement than when one rater 'agrees' and another 'strongly agrees'.

The first step in computing a weighted kappa is to assign weights representing the different levels of agreement for each cell in the KxK contingency table. The weights in the diagonal cells are all 1 (i.e., w_{ii}=1, for all i), and the weights in the off-diagonal cells range from 0 to <1 (i.e., $0 \le w_{ij} < 1$, for all $i \ne j$). These weights are then added to equation (2) to generate a weighted kappa that accounts for varying degrees of agreement or disagreement between the ranked categories:

$$k_w = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}}{1 - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}}$$

The weighted kappa is computed by replacing the probabilities with their respective sample proportions, p_{ij}, p_{i+}, and p_{+i}. If w_{ij}=0 for all $i \ne j$, the weighted kappa coefficient K_w reduces to the standard kappa in equation (2). Note that for binary rating scales, there is no weighted version of kappa, since κ remains the same regardless of the weights used. Again, we can use asymptotic theory and exact methods to estimate confidence intervals and make inferences.

In theory, any weights satisfying the two defining conditions (i.e., weights in diagonal cells=1 and weights in off-diagonal cells ≥0 and <1) may be used. In practice, however, additional constraints are often imposed to make the weights more interpretable and meaningful. For example, since the degree of disagreement (agreement) is often a function of the difference between the ith and jth rating categories, weights are typically set to reflect adjacency between rating categories, such as by $w_{ij}=f(i-j)$, where f is some decreasing function satisfying three conditions: (a) $0 \le f(x) < 1$; (b) $f(x)=f(-x)$; and (c) $f(0)=1$. Based on these conditions, larger weights (i.e., closer to 1) are used for weights of pairs of categories that are closer to each other and smaller weights (i.e., closer to 0) are used for weights of pairs of categories that are more distant from each other.

Two such weighting systems based on column scores are commonly employed. Suppose the column scores are ordered, say $C_1 \le C_2 \ldots \le C_r$ and assigned values of 0,1,…r. Then, the Cicchetti–Allison weight and the Fleiss–Cohen weight in each cell of the KxK contingency table are computed as follows:

Cicchetti-Allison weights: $w_{ij} = 1 - \frac{|C_i - C_j|}{|C_1 - C_r|}$

Fleiss-Cohen weights: $w_{ij} = 1 - \frac{(C_i - C_j)^2}{(C_1 - C_r)^2}$

Example 2. *If depression is categorized into three ranked levels as shown in Table 5, the agreement of the classification based on information provided by the probands with the classification based on information provided by other informants can be estimated using the unweighted kappa coefficient as follows:*

$$k = \frac{\frac{66}{200} + \frac{16}{200} + \frac{27}{200} - \left(\frac{116}{200}\frac{85}{200} + \frac{41}{200}\frac{62}{200} + \frac{43}{200}\frac{53}{200}\right)}{1 - \left(\frac{116}{200}\frac{85}{200} + \frac{41}{200}\frac{62}{200} + \frac{43}{200}\frac{53}{200}\right)} = 0.2812.$$

Applying the Cicchetti-Allison weights (shown in Table 5) to the unweighted formula generates a weighed kappa:

$$k_w = \frac{\left(\frac{66}{200} + \frac{16}{200} + \frac{27}{200}\right) + 0.5 * \left(\frac{13}{200} + \frac{36}{200} + \frac{10}{200} + \frac{12}{200}\right) - \left(\frac{116}{200}\frac{85}{200} + \frac{41}{200}\frac{62}{200} + \frac{43}{200}\frac{53}{200}\right) + 0.5 * \left(\frac{85}{200}\frac{41}{200} + \frac{62}{200}\frac{116}{200} + \frac{62}{200}\frac{43}{200} + \frac{53}{200}\frac{41}{200}\right)}{1 - \left(\frac{116}{200}\frac{85}{200} + \frac{41}{200}\frac{62}{200} + \frac{43}{200}\frac{53}{200}\right) + 0.5 * \left(\frac{85}{200}\frac{41}{200} + \frac{62}{200}\frac{116}{200} + \frac{62}{200}\frac{43}{200} + \frac{53}{200}\frac{41}{200}\right)} = 0.3679.$$

Applying the Fleiss-Cohen weights (shown in Table 5) involves replacing the 0.5 weight in the above equation with 0.75 and results in a K_w of 0.4482. Thus the weighted kappa coefficients have larger absolute values than the unweighted kappa coefficients. The overall result indicates only fair to moderate agreement between the two methods of classifying the level of depression. As seen in Table 5, the low agreement is partly due to the fact that a large number of subjects classified as minor depression based on information from the proband were not identified using information from other informants.

6. Statistical Software

Several statistical software packages including SAS, SPSS, and STATA can compute kappa coefficients. But agreement data conceptually result in square tables with entries in all cells, so most software packages will not compute kappa if the agreement table is non-square, which can occur if one or both raters do not use all the rating categories when rating subjects because of biases or small samples.

Table 5. Three ranked levels of depression categorized based on information from the probands themselves or on information from other informants about the probands

Probands	Other informants			total
	no depression	minor depression	major depression	
no depression	66 (1.0/1.0)[a]	13 (0.5/0.75)[a]	6 (0.0/0.0)[a]	85
minor depression	36 (0.5/0.75)[a]	16 (1.0/1.0)[a]	10 (0.5/0.75)[a]	62
major depression	14 (0.0/0.0)[a]	12 (0.5/0.75)[a]	27 (1.0/1.0)[a]	53
total	116	41	43	200

[a] values in parentheses are the Cicchetti-Allison and Fleiss-Cohen weights used when computing weighted kappa

In some special circumstances the software packages will compute incorrect kappa coefficients if a square agreement table is generated despite the failure of both raters to use all rating categories. For example, suppose a scale for rater agreement has three categories, A, B, and C. If one rater only uses categories B and C, and the other only uses categories A and B, this could result in a square agreement table such as that shown in Table 6. This is a square table, but the rating categories in the rows are completely different from those represented by the column. Clearly, kappa values generated using this table would not provide the desired assessment of rater agreement. To deal with this problem the analyst must add zero counts for the rating categories not endorsed by the raters to create a square table with the right rating categories, as shown in Table 7.

Table 6. Hypothetical example of incorrect agreement table that can occur when two raters on a three-level scale each only use 2 of the 3 levels

Classification of rater 1	Classification of rater 2		total
	B	C	
A	16	2	18
B	5	14	19
total	21	16	37

Table 7. Adjustment of the agreement table (by adding zero cells) needed when two raters on a three-level scale each only use 2 of the 3 levels

Classification of rater 1	Classification of rater 2			total
	A	B	C	
A	0	16	2	18
B	0	5	14	19
C	0	0	0	0
total	0	21	16	37

6.1 SAS

In SAS, one may use PROC FREQ and specify the corresponding two-way table with the "AGREE" option. Here are the sample codes for Example 2 using PROC FREQ:

PROC FREQ DATA = (the data set for the depression diagnosis study);

*TABLE (variable on result using proband) * (variable on result using other informants)/ AGREE;*

RUN;

PROC FREQ uses Cicchetti-Allison weights by default. One can specify (WT=FC) with the AGREE option to request weighted kappa coefficients based on Fleiss-Cohen weights. It is important to check the order of the levels and weights used in computing weighted kappa. SAS calculates weights for weighted kappa based on unformatted values; if the variable of interest is not coded this way, one can either recode the variable or use a format statement and specify the "ORDER = FORMATTED" option. Also note that data for contingency tables are often recorded as aggregated data. For example, 10 subjects with the rating 'A' from the first rater and the rating 'B' from the second rater may be combined into one observation with a frequency variable of value 10. In such cases a weight statement "weight (the frequency variable);" may be applied to specify the frequency variable.

6.2 SPSS

In SPSS, kappa coefficients can be only be computed when there are only two levels in the rating scale so it is not possible to compute weighted kappa coefficients. For a two-level rating scale such as that described in Example 1, one may use the following syntax to compute the kappa coefficient:

CROSSTABS

/TABLES=(variable on result using proband) BY (variable on result using other informants)

/STATISTICS=KAPPA.

An alternatively easier approach is to select appropriate options in the SPSS menu:

1. Click on Analyze, then Descriptive Statistics, then Crosstabs.
2. Choose the variables for the row and column variables in the pop-up window for the crosstab.
3. Click on Statistics and select the kappa checkbox.
4. Click Continue or OK to generate the output for the kappa coefficient.

7. Discussion

In this paper we introduced the use of Cohen's kappa coefficient to assess between-rater agreement, which has the desirable property of correcting for chance agreement. We focused on cross-sectional studies for two raters, but extensions to longitudinal studies with missing values and to studies that use more than two raters are also available.[6] Cohen's kappa generally works well, but in some specific situations it may not accurately reflect the true level of agreement between raters.[7]. For example, when both raters report a very high prevalence of the condition of interest (as in the hypothetical example shown in Table 2), some of the overlap in their diagnoses may reflect their common knowledge about the disease in the population being rated. This should be considered 'true' agreement, but it is attributed to chance agreement (i.e., kappa=0). Despite such limitations, the kappa coefficient is an informative measure of agreement in most circumstances that is widely used in clinical research.

Cohen's kappa can only be applied to categorical ratings. When ratings are on a continuous scale, Lin's concordance correlation coefficient[8] is an appropriate measure of agreement between two raters,[8] and the intraclass correlation coefficients[9] is an appropriate measure of agreement between multiple raters.

Conflict of interest

The authors declare no conflict of interest.

Funding

None.

Kappa 系数：一种衡量评估者间一致性的常用方法

唐万，胡俊，张晖，吴攀，贺华

概述：在精神卫生和社会心理学研究中，常常需要报告研究使用某一评估方法的评估者间的一致性。本文讨论了一致性的概念，强调一致性与相关性的本质区别。Kappa 系数是衡量一致性的一个常用统计方法。我们用几个例子说明如何通过手工计算或统计软件包 SAS、SPSS 等计算 Kappa 系数，用真实的研究数据说明如何在临床研究和实践中使用和解释这个系数。最后文章讨论了该系数的局限性。

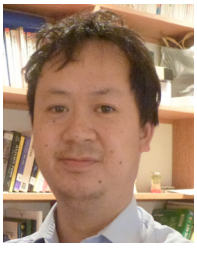
关键词：评估者间一致性，Kappa 系数，加权 Kappa，相关性

本文全文中文版从 2015 年 03 月 25 日起在 www.shanghaiarchivesofpsychiatry.org/cn 可供免费阅览下载

References

1. Spitzer RL, Gibbon M, Williams JBW. *Structured Clinical Interview for Axis I DSM-IV Disorders*. Biometrics Research Department: New York State Psychiatric Institute; 1994
2. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960; **20**(1): 37-46
3. Duberstein PR, Ma Y, Chapman BP, Conwell Y, McGriff J, Coyne JC, et al. Detection of depression in older adults by family and friends: distinguishing mood disorder signals from the noise of personality and everyday life. *Int Psychogeriatr.* 2011; **23**(4): 634-643. doi: http://dx.doi.org/10.1017/S1041610210001808
4. Tang W, He H, Tu XM. *Applied Categorical and Count Data Analysis*. Chapman & Hall/CRC; 2012
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977; **33**: 159-174. doi: http://dx.doi.org/10.2307/2529310
6. Ma Y, Tang W, Feng C, Tu XM. Inference for kappas for longitudinal study data: applications to sexual health research. *Biometrics.* 2008; **64**: 781-789. doi: http://dx.doi.org/10.1111/j.1541-0420.2007.00934.x
7. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol*. 1990; **43**(6): 543-549. doi: http://dx.doi.org/10.1016/0895-4356(90)90158-L
8. Lin L. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989; **45**(1): 255-268. doi: http://dx.doi.org/10.2307/2532051
9. Shrout PE, Fleiss J. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*. 1979; **86**(2): 420-428

(received, 2015-01-28; accepted, 2015-02-04)



Dr. Tang is a Research Associate Professor of Biostatistics in the Department of Biostatistics at the University of Rochester. His research interests are in semi-parametric modeling of longitudinal data with missing values, smoothing methods, and categorical and count data analysis and applications of statistical methods to psychosocial research. Dr. Tang received his PhD in Mathematics from the Department of Mathematics at the University of Rochester in 2004.