•Biostatistics in psychiatry (32)•

Correlation and agreement: overview and clarification of competing concepts and measures

Jinyuan LIU[1], Wan TANG[2], Guanqin CHEN[1], Yin LU[3,4], Changyong FENG[1], Xin M. TU[1,*]

Summary: Agreement and correlation are widely-used concepts that assess the association between variables. Although similar and related, they represent completely different notions of association. Assessing agreement between variables assumes that the variables measure the same construct, while correlation of variables can be assessed for variables that measure completely different constructs. This conceptual difference requires the use of different statistical methods, and when assessing agreement or correlation, the statistical method may vary depending on the distribution of the data and the interest of the investigator. For example, the Pearson correlation, a popular measure of correlation between continuous variables, is only informative when applied to variables that have linear relationships; it may be non-informative or even misleading when applied to variables that are not linearly related. Likewise, the intraclass correlation, a popular measure of agreement between continuous variables, may not provide sufficient information for investigators if the nature of poor agreement is of interest. This report reviews the concepts of agreement and correlation and discusses differences in the application of several commonly used measures.

Keywords: concordance correlation; intraclass correlation; Kendall's tau; non-linear association; Pearson's correlation; Spearman's rho

[*Shanghai Arch Psychiatry.* 2016; **28**(2): 115-120. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.216045]

1. Introduction

Agreement and correlation are widely used concepts in the medical literature. Both are used to indicate the strength of association between variables of interest, but they are conceptually distinct and, thus, require the use of different statistics.

Correlation focuses on the association of changes in two outcomes, outcomes that often measure quite different constructs such as cancer and depression. The Pearson correlation is the most popular measure of the association between two continuous outcomes, but it is only useful when measuring linear relationships between variables. If the relationship is non-linear, the Pearson correlation generally does not provide a good indication of association between the variables. Another problem is that using the standard interpretation of Pearson correlation coefficients can, in some circumstances, lead to incorrect conclusions.

Agreement, also known as reproducibility, is a concept closely related to, but fundamentally different

[1] Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA

[2] Department of Biostatistics & Bioinformatics, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, USA

[3] VA Cooperative Studies Program Palo Alto Coordinating Center, VA Palo Alto Health Care System, Palo Alto, CA, USA

[4] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

*correspondence: Professor Xin M. Tu, Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Ave. Box 630, CTSB 4.239, Rochester, NY 14642, USA.. E-mail: Xin_Tu@URMC.Rochester.edu

A full-text Chinese translation of this article will be available at http://dx.doi.org/10.11919/j.issn.1002-0829.216045 on August 25, 2016.

from, correlation. Like correlation, agreement also assesses the relationships between outcomes of interest, but, as the name indicates, the emphasis is on the degree of concordance in the opinions between two or more individuals or in the results between two or more assessments of the variable of interest. An example of agreement in mental health research is the consensus between multiple clinicians about the psychiatric diagnoses of a group of patients. In biomedical sciences agreement can also include measures of the reproducibility (i.e., reliability) of a laboratory test result when repeated in the same center or when conducted in multiple centers under the same conditions. It is not sensible to speak of agreement (reproducibility) between variables that measure different constructs; so when measuring the association between different variables – such as weight and height – one can assess correlation but not agreement. For continuous outcomes, the intraclass correlation (ICC) is a popular measure of agreement. Like the Pearson correlation, the ICC is an estimate of the magnitude of the relationship between variables (in this case, between multiple assessments of the same variable). However, the ICC also takes into account rater bias, the element that distinguishes agreement from correlation; that is, good agreement (reproducibility) not only requires good correlation, it also requires small rater bias.

In this report, we provide an overview of popular measures and statistical methods for assessing the two different notations of association between variables. We also clarify the key differences between the measures and between the methods used to assess the measures. We focus on continuous outcomes and assume all variables are continuous unless stated otherwise.

2. Correlation measures

2.1 Pearson correlation

Consider a sample of n subjects and a bivariate continuous outcome, (u_i, v_i), from each subject within the sample ($1 \le i \le n$). The Pearson correlation is the most popular statistic for measuring the association between the two variables u_i and v_i:[1]

$$\widehat{p} = \frac{\sum_{i=1}^{n} (u_i - \overline{u}.)\ (v_i - \overline{v}.)}{\sqrt{\sum_{i=1}^{n} (u_i - \overline{u}.)^2}\ \sqrt{\sum_{i=1}^{n} (v_i - \overline{v}.)^2}}, \qquad (1)$$

$$\overline{u}. = \frac{1}{n} \sum_{i=1}^{n} u_i, \quad \overline{v}. = \frac{1}{n} \sum_{i=1}^{n} v_i,$$

where $\overline{u}.(\overline{v}.)$ denotes the sample mean of $u_i(v_i)$ The Pearson correlation $\widehat{p}$ ranges between -1 and 1, with 1(-1) indicating perfect positive (negative) correlation and 0 indicating no association between the variables.

As popular as it is, the Pearson correlation is only appropriate for measuring correlation between u_i and v_i when the two variables follow a linear relationship. If the bivariate outcome (u_i, v_i) follows a non-linear relationship, $\widehat{p}$ is not an informative measure and is difficult to interpret.

To see this, let $\mu_u(\mu_v)$ and $\sigma_u^2(\sigma_v^2)$ denote the (population) mean and (population) variance of the variable $u_i(v_i)$. The Pearson correlation is an estimate of the following product moment correlation:

$$p = Corr(u_i, v_i) = \frac{Cov(u_i, v_i)}{\sqrt{Var(u_i)\, Var(v_i)}} = \left[\frac{E(u_i - \mu_u)(v_i - \mu_v)}{\sqrt{\sigma_u^2 \sigma_v^2}}\right]. \qquad (2)$$

Unlike $\widehat{p}$, which measures correlation between u_i and v_i based on the sample, the product-moment correlation p is the population-level correlation, which cannot be calculated but is estimated by $\widehat{p}$. Thus, $\widehat{p}$ may also be referred to as the 'sample product-moment correlation'.

If u_i and v_i have a linear relationship, then $u_i = av_i + b + \varepsilon_i$, where a and b are some constants, and ε_i denotes random errors with mean 0 and variance σ_ε^2. By centering $u_i(v_i)$ at its mean, we have: $u_i - \mu_u = a(v - \mu_v) + \varepsilon_i$. It follows that $\sigma_u^2 = a^2\sigma_v^2 + \sigma_\varepsilon^2$. If u_i and v_i are perfectly correlated, that is, $\sigma_\varepsilon^2 = 0$, it follows from Equation (2) that p=1 or (-1), depending on whether a is positive or negative. Also, if u_i and v_i are uncorrelated, or independent, that is, a=0, then p=0 and vice versa.

If u_i and v_i have a non-linear relationship, the product moment correlation generally does not provide an informative measure of correlation. The example below shows that the Pearson correlation in this case can be quite misleading.

Example 1. Suppose that u_i and v_i are perfectly correlated and follow the non-linear relationship, $u_i = v_i^9$. Further, assume that v_i follows a standard normal distribution $N(0,1)$ with mean 0 and variance 1. Then, the product-moment correlation is:

$$p = \frac{E(v_i^{10}) - E(v_i^9)E(v_i)}{\sqrt{Var(v_i^9)\, Var(v_i)}} = \frac{E(v_i^{10})}{\sqrt{E(v_i^{18}) - E^2(v_i^9)}} = \frac{E(v_i^{10})}{\sqrt{E(v_i^{18})}} = 0.161. \qquad (3)$$

The poor association between u_i and v_i as indicated by the product-moment correlation contradicts the conceptual perfect correlation between the two variables. Thus, the product-moment and its sample counterpart, the Pearson correlation, generally do not apply to non-linear relationships.

2.2 Spearman's Rho

Spearman's rho is also a popular measure of association. Unlike the Pearson correlation, it also applies to non-linear relationship, thereby addressing the aforementioned limitation associated with the Pearson correlation.

Let $q_i(r_i)$ denote the rankings of $u_i(v_i)$, ($1 \le i \le n$). Spearman's rho is defined as:

$$\widehat{\rho} = \frac{\sum_{i=1}^{n} (q_i - \overline{q}.) \quad (r_i - \overline{r}.)}{\sqrt{\sum_{i=1}^{n} (q_i - \overline{q}.)^2 \quad \sum_{i=1}^{n} (r_i - \overline{r}.)^2}}, \tag{4}$$
$$\overline{q}. = \frac{1}{n}\sum_{i=1}^{n} q_i, \quad \overline{r}. = \frac{1}{n}\sum_{i=1}^{n} r_i.$$

By comparing (1) and (4), it is clear that $\widehat{\rho}$ is really the Pearson correlation when applied to the rankings (q_i, r_i) of the original variables (u_i,v_i). Since the rankings only concern the ordering of the observations, relationships among the rankings are always linear, regardless of whether the original variables are linearly related. Thus, Spearman's rho not only has the same interpretation as the Pearson correlation, but also applies to non-linear relationships.

The Spearman $\widehat{\rho}$ ranges between -1 and 1, with 1 and -1 indicating perfect positive (negative) correlation; when $\widehat{\rho}$=0 there is no association between the variables u_i and v_i. If $\widehat{\rho}$ =1 then $q_i = r_i$, in which case,

$$u_i<u_j,\ v_i<v_j \text{ or } u_i>u_j,\ v_i>v_j \text{ for all } 1\leq i<j\leq n. \tag{5}$$

If $\widehat{\rho}$=-1, then q_i=n-r_i+1, in which case,

$$u_i<u_j,\ v_i>v_j \text{ or } u_i>u_j,\ v_i<v_j \text{ for all } 1\leq i<j\leq n. \tag{6}$$

Any two pairs of bivariate outcomes (u_i,v_i) and (u_j,v_j) that satisfy (5) or (6) are said to be concordant or discordant; that is, u_i and v_i are either both larger or both smaller than u_j and v_j. Thus, perfect positive (negative) correlation by Spearman' rho corresponds to perfect concordance (discordance); that is, concordant (discordant) pairs (u_i,v_i) and (u_j,v_j) for all $1\leq i<j\leq n$.

Example 2. Table 1 shows 12 observations of the bivariate outcome (u_i,v_i) as described in Example 1, and the ranks associated with these observations. Note that u_i and v_i are perfectly related, so their rankings are identical; that is, $q_i = r_i$.

In this example the Pearson correlation $\widehat{\rho}$=0.531, while Spearman's $\widehat{\rho}$ =1. Thus, only the Spearman rho captures the perfect non-linear relationship between u_i and v_i.

Note that the Pearson correlation $\widehat{\rho}$=0.531 has a higher upward bias than the product-moment correlation p=0.161; this occurs due to the small sample size, n=12. As sample size increases, $\widehat{\rho}$ becomes closer to p, a property known as 'consistency' in statistics. For example, we also simulated (u_i,v_i) with n=1000 and obtained $\widehat{\rho}$ =0.173, much closer to p.

Like the Pearson correlation, the Spearman's rho in (4) is a statistic based on a sample. This sample Spearman rho is an estimate of the following population Spearman rho:

$$\rho=12E[I(u_j<u_i)I(v_k<v_i)]-3, \text{ for all } 1\leq i<j<k\leq n. \tag{7}$$

In Equation (7), $E[I(u_j<u_i)I(v_k<v_i)]$ stands for the mathematical expectation of $I(u_j<u_i)I(v_k<v_i)$ and $I(u_j<u_i)$ (similarly $I(v_k<v_i)$) denotes an indicator with $I(u_j<u_i)$=1(0) if $u_j<u_i$. It can be shown that $\widehat{\rho}$ =1(-1) if (u_i,v_i) are perfectly concordant (discordant) and vice versa.

Note that the sample Spearman's rho in (4) is referred to as Spearman's rho in the literature. Unlike the Pearson correlation, there is no formal name for the population Spearman's rho in (7). In general, the lack of a formal name for the population version does not cause confusion, since it is usually clear which one is used within the context of a discussion. Like all statistics, the population version of a statistic is called a parameter in statistical lingo. The statistic and parameter serve different purposes. For example, only the parameter can be used in stating statistical hypotheses, such as the null hypothesis, H:ρ=0, for testing whether the population Spearman's rho is 0. Reported values of Spearman's rho by studies are always the sample Spearman rho.

2.3 Kendall's Tau

Another alternative for non-linear association is Kendall's tau.[2] Like Spearman's rho, Kendall's tau also exploits the concept of concordance and discordance to derive a measure for bivariate outcomes. Unlike Spearman's rho, it uses the notion of concordant and discordant pairs directly in the definition of this correlation measure.

Specifically, Kendall's τ (sample version) is defined as:

$$\begin{aligned} \widehat{\tau} &= \frac{n_c - n_d}{n_t}, \\ n_t &= \frac{1}{2}n(n-1), \\ n_c &= \text{number of concordant pairs}, \\ n_d &= \text{number of discordant pairs}. \end{aligned} \tag{8}$$

In the above, $n_t = \frac{1}{2}n(n-1)$ is the total number of concordant and discordant pairs in the sample. If n_c=n_t(n_d=n_t), then $\widehat{\tau}$=1(-1) and vice versa. Also, if there is no association between u_i and v_i, then n_c and n_d should be close to each other and $\widehat{\tau}$ should be close to 0 (not exactly 0 due to sampling variability).

Table 1. A sample of 12 bivariate outcomes (u_i,v_i) simulated with $u_i = v_i^9$ and v_i from standard normal N(0,1).

u_i	0.26	1.49	1.39	0.65	-0.49	-1.38	1.168	0.87	-0.96	2.15	-0.03	-1.08
v_i	0	38.1	19.4	0.02	-0.002	-18.5	4.06	0.29	-0.68	971.6	0	-2.10
q_i(r_i)	6	11	10	7	4	1	9	8	3	12	5	2

Thus, like Spearman's rho, $\hat{\tau}$ =1(-1) corresponds to perfect concordance (discordance). A value of $\hat{\tau}$ close to 0 indicates weaker or no association between the variables u_i and v_i.

Like the Pearson and Spearman correlation, the sample Kendall's $\hat{\tau}$ in (8) estimates the following population parameter:

$$\tau=2E[I(u_i<u_j)I(v_i<v_j)]-1, \text{ for all } 1\leq i<j\leq n.$$

Like its sample counterpart, τ also ranges between -1 and 1. If (5) holds true for all pairs (u_i,v_i) and (u_j,v_j), then $E[I(u_i<u_j)I(v_i<v_j)]=1$ and $\tau=1$. Likewise, if (6) holds true for all pairs, then $E[I(u_i<u_j)I(v_i<v_j)]=0$ and $\hat{\tau}$ =-1. Thus,τ=1(-1) corresponds to perfect concordance (discordance). Finally, if u_i and v_i are independent, then $E\left[I(u_i<u_j)\,I(v_i<v_j)\right]=\frac{1}{2}$ and $\tau=0$. Thus,$\tau=0$ indicates no association between u_i and v_i, and vice versa.

Example 3. Consider the data in Example 2. The sample Kendall's tau $\hat{\tau}$ =-1. Thus, like Spearman's rho, Kendall's tau also provides a sensible measure of association for non-linearly related variables.

3. Agreement and measures of agreement

Agreement, or reproducibility, is another widely used concept for assessing the relationship among outcomes. As indicated in the Introduction, unlike variables considered in correlation analysis, variables considered for agreement must measure the same construct. Conversely, measures of correlation considered in Section 2 generally do not apply to agreement.

Example 4. Consider two judges who rate each subject from a study of 5 subjects sampled from a population of interest using a scale from 1 to 10. Let u_i and v_i denote the two judges' ratings on the ith subject ($1<i<5$). Suppose that the judges' ratings from the subjects are as follows:

$$(u_i,v_i):(1,6), (2,7), (3,8), (4,9), (5,10).$$

Since u_i and v_i are linearly related, the Pearson correlation can be applied, yielding $\hat{\rho}$=1, indicating perfect correlation. However, the data clearly do not indicate perfect agreement; in fact, the two judges hardly agree with one another.

The poor agreement in this hypothetical example is due to bias in judges' ratings. The mean ratings for the two judges are 3 (for u_i) and 8 (for v_i). Thus, despite the perfect correlation between the ratings, the two judges do not have good agreement because of bias in their ratings of the subjects; either u_i has downward or v_i has upward bias (or both).

The issue of bias does not apply to correlation because the variables considered for correlation generally measure different constructs and, thus, typically have different means. For the Pearson correlation, the sample means $\overline{u}.$ and $\overline{v}.$ are removed from the calculations of the correlation in (1), thus, the Pearson correlation is independent of differences between the (sample) means of the variables being correlated.

3.1 Intraclass correlation

Intraclass correlation (ICC) is a popular measure of agreement for continuous outcomes. Like the Pearson correlation, the ICC requires a linear relationship between the variables. However, it differs from the Pearson correlation in one key respect; the ICC also takes into account differences in the means of the measures being considered. In addition, the ICC can be applied to situations where there are three or more separate raters.

Consider a study with n subjects and assume each subject is rated by a different group of K judges. Let y_{ik} denote the rating of the i^{th} subject by the k^{th} judge ($1\leq i\leq n$, $1\leq k\leq K$). The ICC is defined based on the following linear mixed-effects model:[3-5]

$$y_{ik}=\mu+\beta_i+\varepsilon_{ik}, \; 1\leq k\leq K, \; 1\leq i\leq n, \qquad (9)$$
$$\beta_i \sim N(0, \sigma_\beta^2), \; \varepsilon_{ik} \sim N(0, \sigma^2).$$

In the above model, the fixed effect μ is the (population) mean rating of the study population over all possible K judges from the population of judges; that is, the random effect or latent variable. β_i represents the difference between the mean rating of the i^{th} subject and the mean rating of the study population μ. Thus, the sum $u+\beta_i$ represents the mean rating of the i^{th} subject. The intraclass correlation (ICC) is defined as the variance ratio, $p_{\text{ICC}}=\frac{\sigma_\beta^2}{\sigma_\beta^2+\sigma^2}$, of the variance σ_β^2 of the mean rating of the subjects ($u+\beta_i$) to the total variance consisting of σ_β^2 plus the variance σ^2 of the judges.

If there are only two judges (K=2), then under the linear mixed-effects model in (9) the product-moment correlation between y_{i1} and y_{i2} is the same as the ICC; that is, $Corr(y_{i1},y_{i2})=\frac{\sigma_\beta^2}{\sigma_\beta^2+\sigma^2}$. Moreover, y_{i1} and y_{i2} have the same mean (μ) and variance (σ^2). Thus, in this special case, the ICC is the same as the product-moment correlation ($p_{\text{ICC}}= p$). Note that this result is not a contradiction to the data in Example 4, since u_i and v_i do not have the same mean and thus the linear mixed-effects model in (9) does not apply to the data and the ICC no longer serves its intended purpose in this case. However, since differences in means between judges' ratings decrease the ICC, this agreement index may still be applied in this situation to indicate poorer agreement. Follow-up analyses are necessary to determine whether poor agreement is due to bias or large variability or both between the judges.

Example 5. Consider again Example 4 and let $y_{i1}=u_i$ and $y_{i1}=v_i$. By fitting the model in (9) to the data, we obtain estimates $\hat{\sigma}_\beta^2=0$ and $\hat{\sigma}^2$=9.167. Thus, the (sample) ICC based on the data is $\hat{\rho}_{ICC}$=0, which is quite different from the Pearson correlation. Although the judges' ratings are perfectly correlated, agreement between the judges is extremely poor.

Note that $\widehat{p}_{ICC}$ is not a valid measure of agreement between y_{i1} and y_{i2} for the data in Example 5, since the assumption of a common mean between y_{i1} and y_{i2} is not met by the data. However, it is precisely this assumption that makes $\widehat{p}_{ICC}$ totally different from the Pearson correlation $\widehat{p}$=(1). We may revise the model in (9) to account for the bias in the judges' ratings to consider:

$$y_{ik}=\mu_k+\beta_i+\varepsilon_{ik},\ 1\leq k\leq K,\ 1\leq i\leq n, \qquad (10)$$
$$\beta_i \sim N(0, \sigma_\beta^2),\ \varepsilon_{ik} \sim N(0, \sigma^2),$$

where the added fixed-effect μ_k accounts for the difference between the two judges. By fitting the above model, we obtain estimates $\widehat{\sigma}_\beta^2=1.256$, $\widehat{\sigma}^2=0$, $\widehat{\mu}_1=3$ and $\widehat{\mu}_2=5$. Once accounting for bias, the two judges have perfect agreement. The model in (10) also provides mean ratings $\widehat{\mu}_K$ for the judges. The positive estimate $\widehat{\sigma}_\beta^2$ describes the variability among the subjects. Although the correct model for the data, the ICC calculated from the model in (10) no longer has the interpretation as a measure of agreement. In fact, $\frac{\widehat{\sigma}_\beta^2}{\widehat{\sigma}_\beta^2+\widehat{\sigma}^2}=1$, the same as the Pearson correlation $\widehat{p}=1$ as we have calculated in Example 4.

Note since $p_{\text{ICC}}\geq 0$ we can either reverse code some of the judges' ratings or use a different index, such as the concordance correlation, discussed below.

3.2 Concordance correlation

The concordance correlation (CCC) is another measure of agreement which, unlike the ICC, does not assume a common mean for judges' ratings at the outset, so it can be used to assess both the level of agreement and the level of disagreement. However, a major limitation of the CCC is that it only applies to two judges at a time.

Consider a study with n subjects and assume each subject is rated by a different group of two judges. Let y_{ik} again denote the rating of the i^{th} subject by the k^{th} judge ($1\leq i\leq n$, $1\leq k\leq 2$). Let $\mu_k=E(y_{ik})$ and $\sigma_k^2=Var(y_{ik})$, denoting the mean and variance of y_{ik}, and $\sigma_{12}=Cov(y_{i1}, y_{i2})$, denoting the covariance between y_{i1} and y_{i2}. The CCC is defined as:[6]

$$P_{ccc} = \frac{2\sigma_{12}}{\sigma_1^2+\sigma_2^2+(\mu_1-\mu_2)^2}. \qquad (11)$$

Unlike the ICC, no statistical model is assumed in the definition of p_{CCC}. Further, the two judges can come from two different populations of judges with different means and variances.

The CCC p_{CCC} has a nice decomposition, $p_{\text{CCC}}=pC_b$, where p is the product-moment correlation in (2) and C_b is called the bias correction factor given by:

$$C_b = \frac{2}{\frac{\sigma_1}{\sigma_2}+\frac{\sigma_2}{\sigma_1}+\frac{(\mu_1-\mu_2)^2}{\sigma_1\sigma_2}}. \qquad (12)$$

It can be shown that $p_{\text{CCC}}=1(-1)$ if and only if $p=1(-1)$, $\mu_1=\mu_2$ and $\sigma_1^2=\sigma_2^2$.[6] Thus, $p_{\text{CCC}}=1(-1)$ if and only if $y_{i1}=y_{i2}(y_{i1}=-y_{i2})$, that is, when there is perfect agreement (disagreement). The bias correction factor $C_b(0\leq C_b\leq 1)$ in (12) assesses the level of bias, with smaller C_b indicating larger bias. Thus, unlike the ICC, poor agreement can result from low correlation (small p) or large bias (small C_b).

Example 6. Consider again Example 5. The (sample) mean and variance of y_{i1}, and the (sample) correlation between y_{i1} and y_{i2} are given by: $\widehat{\mu}_1=3$, $\widehat{\mu}_2=8$, $\widehat{\sigma}_1^2=2.5$, $\widehat{\sigma}_2^2=2.5$ and $\widehat{\sigma}_{12}=1$. Thus, it follows from (11) that

$$\widehat{p}_{ccc} = \frac{2\widehat{\sigma}_{12}}{\widehat{\sigma}_1^2+\widehat{\sigma}_2^2+(\widehat{\mu}_1-\widehat{\mu}_2)^2} = 0.053$$

. We can also obtain $\widehat{p}_{CCC}$ by using the decomposition result, which in our case yields $\widehat{p}=1$, $\widehat{C}_b=0.0533$ and $\widehat{p}_{CCC}=\widehat{p}\widehat{C}_b=0.0533$.

Note that unlike correlation the issue of linear versus non-linear association does not arise when assessing agreement. This is because good agreement requires an approximate linear relationship between the outcomes. For example, in the case of two raters, good agreement requires that y_{i1} and y_{i2} are close to each other, such as $y_{i1}=y_{i2}$ in the case of perfect agreement.

4. Discussion

We discussed the concepts of agreement and correlation and described various measures that can be used to assess the relationships among variables of interest. We focused on the measures and methods for continuous outcomes. For non-continuous outcomes, different methods must be applied. For example, for categorical outcomes a different version of Kendall's tau, known as Kendall's tau b can be used for assessing correlation and Kappa can be used for assessing agreement.[7]

Funding

The work was supported in part by a grant (GM108337) from the National Institutes of Health and the National Science Foundation (Tang and Tu) and a pilot grant (UR-CTSI GR500208) from the Clinical and Translational Sciences Institute at the University of Rochester Medical Center (Feng and Tu).

Conflict of interest statement

The authors report no conflict of interest.

Authors' contributions

All authors worked together on this manuscript. In particular, JYL, WT and XMT made major contributions to the section on correlation, GQC, YL and CYF made major contributions to the section on agreement, and JYL and XMT drafted and finalized the manuscript. All authors read and approved the final manuscript.

相关性和一致性：这对相仿概念和测量方法的回顾与阐明

Liu JY, Tang W, Chen GQ, Lu Y, Feng CY, Tu XM

概述：一致性 (agreement) 和相关性 (correlation) 是两个广泛使用的概念，用来评估变量之间的关联。虽然二者相似且相关，但是它们代表关联完全不同的概念。评估变量之间的一致性假设变量测量的是相同的结构，而在变量测量完全不同的结构时也可以评估它们之间的相关性。这种概念上的差异就要求使用不同的统计方法，并且当评估一致性或相关性时，统计方法根据数据的分布和研究者的兴趣可能会有所不同。例如，Pearson 相关性，作为评估连续变量之间相关性的一种普遍测量方法，只有用于符合线性关系的变量时才能提供有用的信息；当用于不符合线性关系的变量时就无法提供准确信息甚至会产生误导。同样地，内部相关性，作为一种评估连续变量之间一致性的常用方法，如果一致性不好的实质正好是研究兴趣所在，那么该测量就不能为研究者提供充分的信息。本报告回顾了一致性和相关性的概念，并讨论了几种常用方法在应用中的差异。

关键词：积差相关性，内部一致性，Kendall's tau，非线性相关，Pearson's 相关性，Spearman's rho

本文全文中文版从 2016 年 8 月 25 日起在 http://dx.doi.org/10.11919/j.issn.1002-0829.216045 可供免费阅览下载

References

1. Stigler SM. Francis Galton's Account of the Invention of Correlation. *Statist Sci.* 1989; **4**(2): 73-79. doi: http://dx.doi.org/10.1214/ss/1177012580
2. Kowalski J, Tu XM. *Modern Applied U Statistics.* New York: Wiley; 2007
3. Lu N, Chen T, Wu P, Gunzler D, Zhang H, He H, *et al*. Functional response models for intraclass correlation coefficients. *Applied Statistics.* 2014; **41**: 2539-2556. doi: http://dx.doi.org/10.1080/02664763.2014.920780
4. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods.* 1996; **1**: 30-46. doi: http://dx.doi.org/10.1037/1082-989X.1.4.390
5. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979; **86**(2): 420-428
6. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989; **45**(1): 255-268
7. Tang W, He H, Tu XM. *Applied Categorical and Count Data Analysis.* Boca Raton, FL: Chapman & Hall/CRC; 2012

Ms. Jinyuan Liu obtained her bachelor's of science degree in statistics from Nanjing University of Posts and Telecommunications in 2015. She is currently a master's student in the Department of Biostatistics and Computational Biology at the University of Rochester in New York, USA. Her research interests include categorical data analysis, machine learning, and social networks.