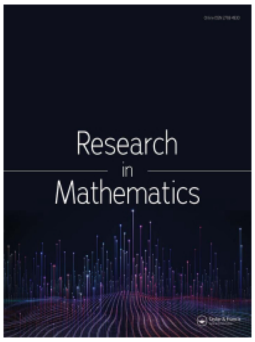


Taylor & Francis
Taylor & Francis Group

Research in Mathematics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/oama23

The behaviour of rank correlation coefficients for incomplete data

Cahyo Crysdian |

To cite this article: Cahyo Crysdian | (2022) The behaviour of rank correlation coefficients for incomplete data, Research in Mathematics, 9:1, 2107793, DOI: 10.1080/27684830.2022.2107793

To link to this article: https://doi.org/10.1080/27684830.2022.2107793

Published online: 07 Aug 2022.

RESEARCH IN MATHEMATICS
2022, VOL. 9, NO. 1, 2107793
https://doi.org/10.1080/27684830.2022.2107793

COMPUTATIONAL SCIENCE | RESEARCH ARTICLE

OPEN ACCESS

The behaviour of rank correlation coefficients for incomplete data

Cahyo Crysdian[a]

[a]Computer Science Department, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia

ABSTRACT

This paper presents the analysis to disclose the behaviour of rank correlation coefficients under the complete and incomplete data condition. The main concern of this research is to deal with the missing data by preserving the originality of data pair rather than experiencing data deletion or imputation. The paper introduces the variability function that is developed for each correlation coefficient in order to disclose the mean and the variance for every possible data sequences. The comparisons between Kendall, Spearman, and the absolute distance measure for index ranking demonstrate the use of variability function under both the complete and incomplete data, in which it becomes a useful tool to describe the coefficient's mechanism to proceed with a set of possible data sequences. The analysis proves that Kendall coefficient becomes the better method compared to Spearman and the absolute distant measure due to threefold, i.e. the ability to preserve the zero mean of variability distribution in complete data, the ability to survive from the missing data, and the ability to gain a higher rate of convergence in incomplete condition. Meanwhile, Spearman fails to preserve the original data pair under the incomplete condition due to direct measurement of rank distances.

ARTICLE HISTORY
Received 13 February 2021
Accepted 27 July 2022

KEYWORDS
Rank correlation coefficient; index ranking; incomplete data; missing data; variability function; variance estimate

Introduction

Rank correlation coefficient aims to measure the association between a pair of ordinal variables representing the ranking of different items obtained from an observation, or the ranking of the same item from different observations. Two famous classical methods that are still widely used even today are Kendall τ and Spearman ρ (Alvo & Cabilio, 1995; Alvo & Park, 2002; Kendall, 1938; Spearman, 1904; Szmidt & Kacprzyk; Xu et al., 2010; Szmidt & Kacprzyk). These methods are formulated by

$$\tau = \frac{nc - nd}{\frac{1}{2}n(n-1)} \tag{1}$$

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{2}$$

in which nc and nd denote the number of concordant and discordant, respectively; d is the distant between a pair of data rank, while n is the number of data. The methods produce a score in a range of $[-1, 1]$ that represents a perfect opposite correlation to a perfect correlation, respectively, while score 0 means that the data pair is independent to each other. Unfortunately, these classical rank correlation coefficients were not designed to deal with the missing data that is often found in many practical observations. As shown by Equations 1 and 2, both Kendall and Spearman are influenced only by a single n number of data. This condition implicitly presents the need to achieve a complete data. Let $(A, B) = (a_1 \cdots a_n, b_1 \cdots b_n)$ become a pair of distinct variables storing the result of an observation. Data completeness is achieved when $a_i \notin \emptyset$ and $b_i \notin \emptyset, i = 1 \cdots n$ with $\emptyset$ denotes the missing data; hence, $|A| = |B| = n$. If somehow, without reducing its generality, $\exists a_i \in \emptyset, |A| < n$ and $b_i \notin \emptyset, i = 1 \cdots n, |B| = n$, thus $|A| \neq |B|$, then the data are incomplete. The incompleteness is mostly due to unmeasured objects in an observation.

However, it is difficult to always achieve data completeness in practical situations due to vary conditions such as hardware and time constraints that restrict object measurements. Therefore, in order to achieve a complete data, the removal of unobserved objects must often be taken prior to applying rank correlation coefficient that includes a list-wise or a pair-wise deletion as noted by Alvo and Cabilio (1995), Alvo and Park (2002), and Raykov et al. (2014). In case of a large number of objects influenced by the deletion process, the data would suffer from significant losses that potentially jeopardize its characteristic. This condition is undesired by most researchers. Therefore, Alvo and

CONTACT Cahyo Crysdian cahyo@ti.uin-malang.ac.id Computer Science, Universitas Islam Negeri Maulana Malik Ibrahim, Indonesia
Reviewing editor: Lishan Liu Qufu Normal University, Shandong, CHINA

Cabilio (1995) introduced the imputation approach based on distance metrics to extend the classical rank correlation coefficient based on Spearman and Kendall to deal with the missing data. The process was developed by associating and relabeling a set of items composing the original incomplete data based on the compatibility between the incomplete and complete data. Later, Cabilio and Tilley (1999) and Alvo and Park (2002) extended this approach based on a multivariate statistical test. Kidwell et al. (2008) even employed Alvo and Cabilio's approach for visualization purposes. Different techniques were presented by Albers and Teulings (1996) to introduce the correlation estimate by incorporating additional information from further observations. This effort increased the size of data to become $n + m_1 + m_2$ with n is the size of the original variable that might contain missing data, while m_1 and m_2 were the size of additional observation obtained from the first and second variables being correlated, respectively. Meanwhile, Raykov et al. (2014) built a correlation estimate by developing a set of predictive rankings to the missing data that utilized the assumption of missing at random. This effort was extended by Eekhout et al. (2015) to include an auxiliary variable in terms of item score information. Recently, various methods have been introduced to predict the missing data such as Kim and Im (2018), Emmanuel et al. (2021), and (Mirzaei et al., 2022) to develop multiple imputation approach, Yan et al. (2021) to predict missing attribute and restore big data by using K-means and Neural Network Backpropagation, and Rejeb et al. () to estimate missing values using Kohonen map.

Despite the progress being made to deal with the incomplete data, the original observed variables, however, become the most appropriate representation to describe system characteristics or phenomenon being investigated, regardless the condition that they might contain some missing data. Altering the original variables means putting into risk on changing or even diminishing data characteristics such as shown by (Zidan et al., 2017,), Kim et al. (2020), Abdel-Aty et al. (2020), and Mirzaei et al., 2022). Conducting imputation for a rank of indices jeopardizes unique features of an index ranking. It is important to note that each index represents different entity associated with an object in a sequence. Therefore, it would not be appropriate to replace an index with its neighbors or to modify an index ranking, since the action would cause the changes in the original variable. The last statement becomes the foundation of this study. Here, we assume that an entity corresponds to only an index, and there should be no repeated index in a ranking. Hence, any sequence of data in this study is recognized as an index ranking.

Meanwhile, a measure of similarity based on the absolute distance between a pair of index ranking $(A, B) = (a_1 \cdots a_n, b_1 \cdots b_m)$ is designed for incomplete data (Crysdian, 2018) as formulated by

$$c = \frac{1}{\min(n, m)} \sum_{i=1, j=1}^{n,m} \frac{\alpha}{|i - j| + 1} \tag{3}$$

with

$$\alpha = \begin{cases} 1 & \text{if } a_i = b_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Equation 3 produces a range of score $[0, 1]$ that presents the independent to tightly correlated data, respectively. This approach enables the similarity measure between a pair of index rankings in whatever condition they might have, and therefore it preserves the originality of the observed variables. For the case of a complete data in which $n = m$, reaching $c = 0$ is not possible due to $a_i \in B$ and $b_j \in A$. While different condition is found from the incomplete data in which n, Equation 3 is capable of reaching $c = 0$ due to the possibility that $a_i \notin B$ or $b_j \notin A$.

From this point, the difference behaviour between varied rank correlation coefficients, i.e. Kendall, Spearman, and the absolute distance measure, are noticeable. As noted by Xu et al. (2010), the problem of mathematical tractability is raised from this issue due to the function complexity to describe the unique mechanism of each coefficient. Hence, it is interesting to reveal how these coefficients behave for $n \to \infty$ under the complete and incomplete conditions. The study aims to discuss the characteristics of rank correlation coefficient to adapt with the complete and incomplete data by disclosing their behaviour for $n \to \infty$ through the application of variability function introduced in Section Material and methods, sub section Coefficient's behaviour for complete data. Here we show how variability function discloses the internal mechanism of each rank correlation coefficient. The rest of the paper is organized as follows. Section 2 discusses the complexity of complete and incomplete condition, and to introduce the variability function to analyze the behaviour of rank correlation coefficient by presenting the following materials, i.e. the complexity of data sequence under complete and incomplete condition in Sub-Section 2.1, the behaviour of each rank correlation coefficient in the complete data by formulating the variability function in Sub-Section 2.2, and the extension of variability function to adapt with the incomplete condition in Sub-Section 2.3. Section 3 discusses the characteristic of rank correlation coefficient by disclosing their behaviours under the condition of complete and incomplete data through the application of variability function. The study is concluded in Section 4.

Material and methods

The complexity of complete and incomplete data

For the case of a complete data $(A, B) = (a_1 \cdots a_n, b_1 \cdots b_n)$, the correlation between A and B is a permutation of n data items. Hence, the correlation score is distributed over a number of possible data sequence N as defined by

$$N(n) = n! \tag{5}$$

Different condition is found for incomplete data $(A, B) = (a_1 \cdots a_n, b_1 \cdots b_m)$, in which n. Without reducing any generality, this assumption states that the incompleteness is suffered by B. Therefore, the possible data sequence of B is the permutation of the result of a union operation between the power set of A with the missing data that is represented by "0". We use "0" for the notation of missing data since the data rank is a positive integer, while the number of "0" in B represent the number of missing data. For instance, $A = \{1, 2, 3\}$; hence, B can be in any permutation of the following combinations, i.e. $\{1, 2, 3\}$, $\{1, 2, 0\}$, $\{1, 3, 0\}$, $\{2, 3, 0\}$, $\{1, 0, 0\}$, $\{2, 0, 0\}$, $\{3, 0, 0\}$, and $\{0, 0, 0\}$. For $B = \{1, 0, 0\}$, which contains two missing data; hence, we can rewrite to become $B = \{1, 0_1, 0_2\}$. There are 3! combination of data sequence that consists of $\{1, 0_1, 0_2\}$, $\{1, 0_2, 0_1\}$, $\{0_1, 1, 0_2\}$, $\{0_2, 1, 0_1\}$, $\{0_1, 0_2, 1\}$, $\{0_2, 0_1, 1\}$. Possible sequence of data grows to become

$$N(n) = (2^n - 1)n! \; + 1 \tag{6}$$

It is important to note that the statement "+ 1" in the last fraction of Equation 6 is to accommodate $\{0, 0, 0\}$. Here, we prefer to exclude it and compute only the possible sequence that carries component from the original data. Hence, total possible sequence of data can be reduced to become $N(n) = (2^n - 1)n!$. Moreover, it is possible to further reduce the possible sequence of data by removing the permutations from the combinations having repeated items, such as $\{1, 0, 0\}$ that comes from $\{1, 0_1, 0_2\}$ and $\{1, 0_2, 0_1\}$. Then, the total possible sequence of data in Equation 6 is reduced to become

$$N(n) = \sum_{i=1}^{n} C_{n,i} P_{n,i} \tag{7}$$

Table 1. Number of possible data sequence

n	Complete data	Incomplete data
1	1	1
2	2	6
3	6	33
4	24	208
5	120	1545
6	720	13,326
7	5040	130,921
8	40,320	1,441,728
9	362,880	17,572,113
10	3,628,800	234,662,230
⋮	⋮	⋮
n	$N(n) = n!$	$N(n) = \sum_{i=1}^{n} C_{n,i} P_{n,i}$

with $C_{n,i} = \frac{n!}{(n-i)!i!}$ and $P_{n,i} = \frac{n!}{(n-i)!}$

The comparison between the number of data sequences for both complete and incomplete data for any n as given in Table 1 shows the exponential growth of possible data sequences. It is difficult to obtain the mean and variance from rank correlation coefficients for large n since the computation involves a large number of possible data sequences. This problem restricts the effort to present the function behavior of rank correlation coefficients, particularly under incomplete condition. Therefore, it is crucial to define the variability function that carries the smallest component of rank correlation coefficient. Then, the computation to obtain the mean and the variance of correlation coefficients can be established using the predefined variability function as elaborated in the next sub-section.

Coefficient's behaviour for complete data

For a pair of complete data $(A, B) = (a_1 \cdots a_n, b_1 \cdots b_n)$, it is not possible to have a repeated index due to the nature of index ranking as stated previously, hence $a_i \neq a_j$ and $b_i \neq b_j$, $i \neq j$, for $i, j = 1 \cdots n$. Kendall coefficient in Equation 1 is computed based on the number of concordant and discordant as follows:

$$nc = \sum_{i=1, j=1, i \neq j}^{n,n} \mathrm{con}(i, j) \tag{8}$$

$$nd = \sum_{i=1, j=1, i \neq j}^{n,n} \mathrm{dis}(i, j) \tag{9}$$

with

$$\mathrm{con}(i, j) = \begin{cases} 1 & \text{if} \quad ((a_i > a_j) \text{ and } (b_i > b_j)) \text{ or } ((a_i < a_j) \text{ and } (b_i < b_j)) \\ 0 & \text{if} \quad \text{otherwise} \end{cases} \tag{10}$$

$$\text{dis}(i,j) = \begin{cases} 1 & \text{if} \quad ((a_i > a_j) \text{ and } (b_i < b_j)) \text{ or } ((a_i < a_j) \text{ and } (b_i > b_j)) \\ 0 & \text{if} \quad \text{otherwise} \end{cases} \tag{11}$$

Hence, the variability function of Kendall coefficient v_K for any n is also defined by the number of concordant and discordant as follows:

$$\begin{aligned} v_K &= nc - nd \\ &= \tau\, n(n-1)/2 \end{aligned}$$

with τ is the Kendall's range of score in $[-1, 1]$, hence $-1 \leq \tau \leq 1$, therefore

$$-n(n-1)/2 \leq v_K \leq n(n-1)/2 \tag{13}$$

Thus, the variability function to satisfy Equation 13 is

$$v_K = \frac{1}{2}n(n-1) - 2i \tag{14}$$

or

$$v_K = -\frac{1}{2}n(n-1) + 2i \tag{15}$$

for $i = 0 \cdots n(n-1)/2$.

Proof:

The variability of Kendal's score is governed by $\frac{1}{2}n(n-1)$ for n number of an index ranking as stated by Equation 1 that become the total number of nc and nd, hence

$$nc + nd = \frac{1}{2}n(n-1) \tag{16}$$

Thus, the number of nc varies between a range of $[0, n(n-1)/2]$. The same condition is applied to nd. To compute $v_K = f(nc, nd)$ in Equation 12, we need to state the extreme maximum or the extreme minimum that can be reached through Equation 16, i.e. $\frac{1}{2}n(n-1)$ or $-\frac{1}{2}n(n-1)$, respectively. We can then visit all possible v_K in Equation 13 by using an order sequence of variable i in a range of $[0, n(n-1)/2]$. QED

Based on (14), $\mu_K = 0$ for any n.

Proof:

$$\begin{aligned} \mu_K &= \frac{\sum_{i=1}^{N(n)} \tau_i}{N(n)} \\ &= \frac{1}{n!}\sum_{i=1}^{n!} \frac{nc_i - nd_i}{n(n-1)/2} \\ &= \frac{2}{n(n-1)n!}\sum_{i=1}^{n!} nc_i - nd_i \end{aligned} \tag{17}$$

Since $\sum_{i=1}^{n!} nc_i - nd_i$ includes all possible sequence that create the variability of Kendal's function as stated in Equation 12, hence

$$\frac{1}{n!}\sum_{i=1}^{n!} nc_i - nd_i \cong \frac{1}{\frac{n(n-1)}{2} + 1} \sum_{i=0}^{n(n-1)/2} v_{Ki} \tag{18}$$

In Equation 18, we include the denominator since it becomes the source of the nominator as stated in Equation 17. Therefore, replacing the nominator has a consequence of replacing the denominator in order to have a fair computation. It is important to note that the "+ 1" statement in the denominator of the right side of Equation 18 represents the sources of i that are started from zero. The mean of Kendall coefficient can then be computed by inserting Equation 18 to Equation 17 as follows:

$$\begin{aligned} \mu_K &= \frac{2}{n(n-1)\left(\frac{n(n-1)}{2} + 1\right)} \sum_{i=0}^{n(n-1)/2} v_{Ki} \\ &= \frac{2}{n(n-1)\left(\frac{n(n-1)}{2} + 1\right)} \sum_{i=0}^{n(n-1)/2} \frac{1}{2}n(n-1) - 2i \\ &= \frac{2}{n(n-1)\left(\frac{n(n-1)}{2} + 1\right)} .0 \\ L &= 0 \quad \text{QED} \end{aligned} \tag{19}$$

Even though the denominator does not influence the computation due to the nominator's symmetry that produces zero mean in Equation 19, the statement of the denominator in Equation 18 is vital for the calculation of the variance. Hence, the variance of Kendall coefficient is obtained by

$$\begin{aligned}\sigma_K^2 &= \frac{1}{N(n)}\sum_{i=1}^{N(n)}(\tau_i-\mu)^2\\ &= \frac{1}{n!}\sum_{i=1}^{n!}\tau_i^2\\ &= \frac{1}{n!}\sum_{i=1}^{n!}\left(\frac{nc_i-nd_i}{\frac{1}{2}n(n-1)}\right)^2\\ &= \frac{4}{n^2(n-1)^2 n!}\sum_{i=1}^{n!}(nc_i-nd_i)^2 \qquad (20)\end{aligned}$$

The variance estimate of Kendall can then be computed in term of n by inserting Equation 18 to Equation 20 as follows:

$$\begin{aligned}\hat{\sigma}_K^2 &\cong \frac{4}{n^2(n-1)^2\left(\frac{n(n-1)}{2}+1\right)}\sum_{i=1}^{n(n-1)/2}v_{Ki}^2\\ &\cong \frac{8}{n^2(n-1)^2(n(n-1)+2)}\sum_{i=0}^{n(n-1)/2}\\ &\left(\frac{1}{2}n(n-1)-2i\right)^2 \qquad (21)\end{aligned}$$

Meanwhile, the variability of Spearman coefficient in Equation 2 can be defined by the total square distance from a pair of index ranking as follows:

$$\begin{aligned}v_S &= \sum_{i=1}^{n!}d_i^2\\ &= \frac{n(n^2-1)(1-\rho)}{6}\end{aligned}$$

Since the score of Spearman is in a range of $[-1,1]$, hence $-1\le\rho\le 1$, therefore

$$0\le v_S\le n(n^2-1)/3 \qquad (23)$$

For any n, the extreme minimum and maximum of v_S in Equation 23 can be visited by using an order sequence of variable i in a range of $[0\cdots n(n^2-1)/3]$. Thus, the variability function to satisfy Equation 23 is

$$v_S = i \qquad (24)$$

for $i=0\cdots n(n^2-1)/3$.

Based on Equation 24, $\mu_S=0$ for any n.

Proof:

$$\begin{aligned}\mu_S &= \frac{1}{n!}\sum_{i=1}^{n!}\rho_i\\ &= \frac{1}{n!}\sum_{i=1}^{n!}\left(1-\frac{6\sum d_i^2}{n(n^2-1)}\right)\end{aligned}$$

Since $\sum_{i=1}^{n!}\left(1-\frac{6\sum d_i^2}{n(n^2-1)}\right)$ includes all possible sequence that create the variability function of Spearman, hence

$$\frac{1}{n!}\sum_{i=1}^{n!}\left(1-\frac{6\sum d_i^2}{n(n^2-1)}\right)\cong\frac{1}{(n(n^2-1)/3)+1}\sum_{i=0}^{n(n^2-1)/3}\left(1-\frac{6v_S}{n(n^2-1)}\right) \qquad (26)$$

Justification of Equation 26 is similar to Equation 18 for Kendall which includes all possible variability and their sources. The mean of Spearman coefficient is computed by inserting Equation 26 to Equation 25 as follows:

$$\begin{aligned}\mu_S &= \frac{1}{(n(n^2-1)/3)+1}\sum_{i=0}^{n(n^2-1)/3}\left(1-\frac{6i}{n(n^2-1)}\right)\\ &= 0 \quad \text{QED}\end{aligned}$$

The result of Equation 25 is consistent with the symmetry of Spearman score. Hence, Spearman's variance for complete data is defined by

$$\begin{aligned}\sigma_S^2 &= \frac{1}{n!}\sum_{i=1}^{n!}(\rho_i-\mu_S)^2\\ &= \frac{1}{n!}\sum_{1=1}^{n!}\rho_i^2\\ &= \frac{1}{n!}\sum_{i=1}^{n!}\left(1-\frac{6\sum d_i^2}{n(n^2-1)}\right)^2\end{aligned}$$

By inserting Equation 26 to Equation 28, we obtain variance estimate of Spearman coefficient as follows:

$$\hat{\sigma}_S^2\cong\frac{1}{(n(n^2-1)/3)+1}\sum_{i=1}^{n(n^2-1)/3}\left(1-\frac{6i}{n(n^2-1)}\right)^2 \qquad (29)$$

The variability function of the absolute distance measure in Equation 3 is described by the total distance of a pair of index ranking as computed by

$$\begin{aligned}v_c &= \sum_{i,j=1}^{n}\frac{1}{1+|i-j|}\\ &= \frac{e_1}{1}+\frac{e_2}{2}+\cdots+\frac{e_n}{n}\end{aligned}$$

with e_i is the number of the absolute distance obtained by the correlation coefficient for $i = 1 \cdots n$. Right now, we just leave e_i since it is difficult to compute their quantities for a single data pair. However, later on this section, we can obtain the pattern of parameter e for the whole data sequences. Since there exist $c_1 \cdots c_{n!}$ score obtained from $n!$ data sequences for complete data: thus, the mean of correlation score is computed by

$$\begin{aligned}\mu_c &= \frac{\sum_{i=1}^{n!} c_i}{n!} \\ &= \frac{\frac{1}{n}\left(\frac{e_{1,1}}{1} + \frac{e_{2,1}}{2} + \ldots + \frac{e_{n,1}}{n}\right) + \ldots + \frac{1}{n}\left(\frac{e_{1,n!}}{1} + \frac{e_{2,n!}}{2} + \ldots + \frac{e_{n,n!}}{n}\right)}{n!} \\ &= \frac{\frac{\sum_{i=1}^{n!} e_{1,i}}{1} + \frac{\sum_{i=1}^{n!} e_{2,i}}{2} + \ldots + \frac{\sum_{i=1}^{n!} e_{n,i}}{n}}{nn!}\end{aligned} \tag{31}$$

By scrutinizing data sequences for small n, we find that parameter e in Equation 31 are described by

$$\begin{aligned}&\sum_{i=1}^{n!} e_{1,i} = n! = n(n-1)! = nN(n-1) \\ &\sum_{i=1}^{n!} e_{2,i} = 2(n-1)(n-1)! = 2(n-1)N(n-1) \\ &\sum_{i=1}^{n!} e_{3,i} = 2(n-2)(n-1)! = 2(n-2)N(n-1) \\ &\vdots \\ &\sum_{i=1}^{n!} e_{n,i} = 2(n-(n-1))(n-1)! = 2(n-(n-1))N(n-1)\end{aligned} \tag{32}$$

Inserting Equation 32 to Equation 31 delivers the following result

$$\hat{\mu}_c = \frac{1}{n}\left(1 + \frac{2}{n}\sum_{i=1}^{n-1}\frac{n-i}{i+1}\right) \tag{33}$$

We could rewrite Equation 33 into a longer form to disclose the pattern of fraction units composing the mean for each n as follow:

$$\begin{aligned}\hat{\mu}_c &= \frac{1}{n}\left(1 + \frac{2(n-1)}{n(1+1)} + \frac{2(n-2)}{n(2+1)} + \ldots + \frac{2(n-(n-1))}{n((n-1)+1)}\right) \\ &= \frac{1}{n}(X_1 + X_2 + \ldots + X_n)\end{aligned} \tag{34}$$

The variance for large n is then computed by

$$\sigma_c^2 = \frac{\sum_{i=1}^{n!}\left(c_i - \mu_c\right)^2}{n!} \tag{35}$$

In this case, $c_i \cong X_j$ for $i = 1 \cdots n!$, $j = 1 \cdots n$, with $n!$ and n become the sources of c_i and X_j respectively. Hence, the variance estimate is obtained by inserting Equation 34 to Equation 35 as follows:

$$\begin{aligned}\hat{\sigma}_c^2 &= \frac{\sum_{j=1}^{n}(X_j - \mu_c)^2}{n} \\ &= \frac{\left(1-\mu_c\right)^2 + \left(\frac{2(n-1)}{n(1+1)} - \mu_c\right)^2 + \left(\frac{2(n-2)}{n(2+1)} - \mu_c\right)^2 + \cdots + \left(\frac{2(n-(n-1))}{n((n-1)+1)} - \mu_c\right)^2}{n} \\ &= \frac{\left(1-\mu_c\right)^2 + \sum_{i=1}^{n}\left(\frac{2(n-i)}{n(i+1)} - \mu_c\right)^2}{n}\end{aligned} \tag{36}$$

Incomplete condition

In order to adapt with the incomplete data $(A, B) = (a_1 \cdots a_n, b_1 \cdots b_m)$ in which $\exists a_i \notin B$ or $\exists b_j \notin A$, hence n, it is vital to preserve both n and m as the domain of correlation. Altering either n or m by the deletion or adding more observation would make correlation running in a different environment. Even though some researchers or statisticians might be interested to observe the cause of missing data such as the missing completely at random (MCAR) or missing at random (MAR) in order to build the most suitable distribution for the sake of prediction or imputation; however, this issue is beyond the scope of the paper that concerns with the original data pair. Disclosing the survivability of rank correlation coefficient under the incomplete data is more desired. Therefore, the analysis focusses on observing the effect of the missing data based on the mean and the variance for $n \rightarrow \infty$.

For the case of Kendall coefficient, since the measurement is possible to take place only for $a_i \in B$ and $b_j \in A$; hence, it is necessary to introduce α in order to compute nc and nd by excluding $a_i \notin B$ and $b_j \notin A$ as follows:

$$nc = \sum_{i=1}^{k} \alpha\ con(i,j) \tag{37}$$

$$nd = \sum_{i=1}^{k} \alpha\ dis(i,j)$$

with α is defined in Equation 4. Due to Equation 37, the variability function of Kendall coefficient in Equation 12 is expanded to become:

$$\begin{aligned}v_k &= nc - nd \\ &= C_{n-h,2} - 2i \\ &= \frac{1}{2}(n-h)(n-(h+1)) - 2i\end{aligned} \tag{38}$$

with h and i are the range of integers as the function of n in which $h = 0 \dots (n-2)$ and $i = 0 \dots (n-h)(n-(h+1))/2$.

Based on Equation 38, $\mu_K = 0$ for incomplete data.

Proof:

$$\begin{aligned}\mu_K &= \frac{\sum_{i=1}^{N(n)+1} \tau_i}{N(n)+1} \\ &= \frac{2}{n(n-1)(N(n)+1)} \sum_{i=1}^{N(n)+1} nc_i - nd_i \\ &= \frac{2}{n(n-1)(N(n)+1)} \sum_{h=0}^{n-2} \sum_{i=0}^{(n-h)(n-(h+1))/2} v_K \\ &= \frac{2}{n(n-1)(N(n)+1)} \sum_{h=0}^{n-2} \sum_{i=0}^{(n-h)(n-(h+1))/2} \\ & \frac{1}{2}(n-h)(n-(h+1)) - 2i = 0 \\ & QED\end{aligned} \tag{39}$$

Therefore, the variance of Kendall coefficient for incomplete data is described by

$$\begin{aligned}\sigma_K^2 &= \frac{1}{N(n)+1} \sum_{i=1}^{N(n)+1} (\tau_i - \mu)^2 \\ &= \frac{1}{N(n)+1} \sum_{i=1}^{N(n)+1} {\tau_i}^2 \\ &= \frac{4}{n^2(n-1)^2(N(n)+1)} \sum_{i=1}^{N(n)+1} (nc_i - nd_i)^2 \\ &\cong \frac{4}{n^2(n-1)^2 T_v} \sum_{h=0}^{n-2} \sum_{i=0}^{(n-h)(n-(h+1))/2} {v_K}^2\end{aligned} \tag{40}$$

with T_v is the total number of variation generated by v_K in which

$$T_v = \sum_{h=0}^{n-2} \frac{(n-h)(n-(h+1))}{2} \tag{41}$$

Thus, the variance estimate is computed by

$$\begin{aligned}\hat{\sigma}_K^2 = \frac{4}{n^2(n-1)^2 T_V} \sum_{h=0}^{n-2} \sum_{i=0}^{(n-h)(n-(h+1))/2} \\ \left(\frac{1}{2}(n-h)(n-(h+1)) - 2i\right)^2\end{aligned} \tag{42}$$

Meanwhile, it is not possible to define any variability function for Spearman under incomplete data without violating the correlation principle due to undefined distance d either for $a_i \notin B$ or $b_j \notin A$. Referring to v_S in Equation 22, d becomes the core of variability function for Spearman. In this case, defining $d \geq n-1$ as the maximum distance for $a_i \notin B$ or $b_j \notin A$ violates to the range of score $[-1, 1]$, while defining $d = 0$ produces weird measurement since $\rho(\ a_i \notin B\text{or}\ \ b_j \notin A) = \rho(A = B) = 1$. Therefore, Spear man fails to present its variability due to the failure to define the distance for the missing items. Here, we do not argue that Spearman is completely failing to deal with the missing data since any pairwise deletion or data imputation could be employed to achieve data completeness. However, it is vital to obey the consensus stated in the beginning of this section to preserve the originality of data pair. It means taking any action to alter the data either by deletion or imputation violates the foundation of this study.

While the absolute distance measure in Equation 3 is intentionally designed to adapt with the incomplete data, the variability function defined in Equation 30 for complete data complies with the incomplete condition. Hence, it is merely to insert the missing data into its variability function in order to describe its mechanism under the incomplete condition as follows. The mean of this coefficient is computed by

$$\mu_c = \frac{\sum_{i=1}^{N(n)+1} c_i}{N(n)+1} \tag{43}$$

with $N(n)$ is defined in Equation 7. Here, the statement "+1" in the denominator is to accommodate the missing of all items either in A or B. Inserting Equation 30 to Equation 43, we obtain

$$\begin{aligned}\mu_C &= \frac{\sum_{i=1}^{N(n)+1} (v_c/n)_i}{N(n)+1} \\ &= \frac{\frac{\sum_{i=1}^{N(n)+1} e_{1,i}}{1} + \frac{\sum_{i=1}^{N(n)+1} e_{2,i}}{2} + \dots + \frac{\sum_{i=1}^{N(n)+1} e_{n,i}}{n}}{n(N(n)+1)}\end{aligned} \tag{44}$$

To compute Equation 44, we need to modify Equation 32 for incomplete data by inserting the number of possible data sequences in Equation 7. Hence, we obtain

$$\begin{aligned}&\sum_{i=1}^{N_n+1} e_{1,i} = nN(n-1) = n\sum_{i=0}^{n-1} C_{n-1,i}P_{n-1,i} \\ &\sum_{i=1}^{N_n+1} e_{2,i} = 2(n-1)N(n-1) = 2(n-1)\sum_{i=0}^{n-1} C_{n-1,i}P_{n-1,i} \\ &\sum_{i=1}^{N_n+1} e_{3,i} = 2(n-2)N(n-1) = 2(n-2)\sum_{i=0}^{n-1} C_{n-1,i}P_{n-1,i} \\ &\vdots \\ &\sum_{i=1}^{N_n+1} e_{n,i} = 2(n-(n-1))N(n-1) = 2(n-(n-1))\sum_{i=0}^{n-1} C_{n-1,i}P_{n-1,i}\end{aligned} \tag{45}$$

Inserting Equation 45 to Equation 44, we obtain

$$\hat{\mu}_C = \frac{1}{n}\left(n + 2\sum_{i=1}^{n-1}\frac{n-i}{i+1}\right)\left(\frac{\sum_{i=0}^{n-1} C_{n-1,i}P_{n-1,i}}{1+\sum_{i=1}^{n} C_{n-1,i}P_{n-1,i}}\right) \tag{46}$$

To simplify Equation 46, we can write $\beta = \frac{\sum_{i=0}^{n-1} C_{n-1,i}P_{n-1,i}}{1+\sum_{i=1}^{n} C_{n-1,i}P_{n-1,i}}$, hence

$$\begin{aligned}\hat{\mu}_C &= \frac{\beta}{n}\left(n + 2\sum_{i=1}^{n-1}\frac{n-i}{i+1}\right)\\ &= \frac{1}{n}\left(n\beta + \frac{2\beta(n-1)}{1+1} + \frac{2\beta(n-2)}{2+1} + \cdots + \frac{2\beta(n-(n-1))}{(n-1)+1}\right)\\ &= \frac{1}{n}(X_1 + X_2 + \cdots + X_n)\end{aligned} \tag{47}$$

The variance is then computed by

$$\begin{aligned}\sigma_c^2 &= \frac{\sum_{i=1}^{N(n)+1}\left(c_i - \mu_c\right)^2}{N(n)+1}\\ &\cong \frac{\sum_{i=1}^{n}\left(X_i - \mu_c\right)^2}{n}\end{aligned} \tag{48}$$

The variance estimate for the absolute distant-based measure is obtained by inserting Equation 47 to Equation 48 as follows:

Result and discussion

The behaviours of Kendall, Spearman and the absolute distance measure are asymptotically normal under the complete data with $N(0, \hat{\sigma}_K^2)$, $N(0, \hat{\sigma}_S^2)$, and $N(\hat{\mu}_c, \hat{\sigma}_c^2)$, respectively. In this case, $\hat{\sigma}_K^2$ is defined by Equation 21, $\hat{\sigma}_S^2$ is defined by Equation 29, while $\hat{\mu}_c$ and $\hat{\sigma}_c^2$ are defined by Equation 33 and 36. It is worth noting that $\hat{\sigma}_K^2$, $\hat{\sigma}_S^2$, $\hat{\mu}_c$, and $\hat{\sigma}_c^2$ are unbiased since all parameters are perfectly described by their variability function, i.e. v_K for Kendall in both Equation 14 and 15, v_S for Spearman in Equation 24, and v_c for the absolute distance measure in Equation 30, respectively. Graphing the variance estimates of Kendall, Spearman and the absolute distance measure as defined by Equation 21, 29, and 36, respectively, produces Figure 1 that discloses the following phenomenon, i.e. both Kendall and Spearman share similar characteristics for large n, and differ significantly to the absolute distance measure as seen by Figure 1. In this case, Kendall and Spearman directly lead to the convergence, with Spearman has the faster convergence compared to Kendall based on the growth of n. Meanwhile, the absolute distance measure presents a ripple at $n \leq 4$ before leading to the convergence at $n > 4$. Therefore, sorting the rate of convergence from the fastest to the slowest is demonstrated by Spearman, Kendall and the absolute distance measure. The higher convergence rate of Spearman compared to other methods is caused by the larger size of Spearman variability that grows exponentially based on the growth of n as shown by Figure 2. This condition also discloses the slower convergence rate of the absolute distance measure compared to other methods due to the less number of variability based on the growth on n.

Meanwhile, graphing the variability function of each rank correlation coefficient as defined by v_K in both Equation 14 and 15, v_S in Equation 24, and v_c in Equation 30 for Kendall, Spearman, and the absolute distance measure, respectively, reveals the distribution of variability function based on the growth of n as shown by Figure 3A-C. Figure 3A depicts Kendall variability function that perfectly distributes around zero mean. It is different to the variability function of Spearman that distributes above zero and grows exponentially with the n as shown by Figure 3B, and the absolute distance measure that distributes in a range of $[0, 1]$ as shown by Figure 3C. It shows that Kendall is a well-designed method for a rank correlation coefficient that even preserves the zero mean for any n by its variability function.

For the incomplete data, only Kendall and the absolute distance measure survives. It is due to the flexibility to adapt with the missing data by accepting α. It is important to note that the mechanism to employ α is different from the deletion mechanism since the former preserves the originality of data pair as the domain of correlation. In this case, Kendall and the absolute distance measure are asymptotically normal with $N(0, \hat{\sigma}_K^2)$ and $N(\hat{\mu}_c, \hat{\sigma}_c^2)$ respectively, in which $\hat{\sigma}_K^2$, $\hat{\mu}_c$ and $\hat{\sigma}_c^2$ are unbiased as defined by Equation 42, 46, and 49, respectively. Comparing the variance estimates of Kendall against the absolute distance

$$\begin{aligned}\hat{\sigma}_c^2 &= \frac{1}{n}\left(\left(n\beta - \mu_c\right)^2 + \left(\frac{2\beta(n-1)}{2} - \mu_c\right)^2 + \left(\frac{2\beta(n-2)}{3} - \mu_c\right)^2 + \ldots + \left(\frac{2\beta(n-(n-1))}{(n-1)+1} - \mu_c\right)^2\right)\\ &= \frac{1}{n}\left(\left(n\beta - \mu_c\right)^2 + \sum_{i=1}^{n-1}\left(\frac{2\beta(n-i)}{i+1} - \mu_c\right)^2\right) = \frac{1}{n}\left(\left(n\beta - \mu_c\right)^2 + \sum_{i=1}^{n-1}\left(\frac{2\beta(n-i)}{i+1} - \mu_c\right)^2\right)\end{aligned} \tag{49}$$

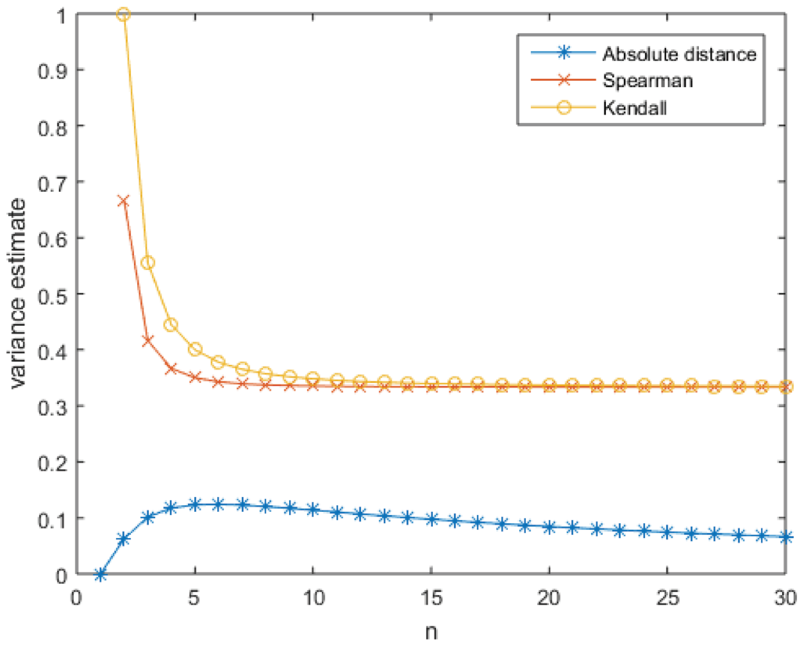


Figure 1. The variance estimates for kendal, Spearman, and the absolute distance measure under complete data condition based on the growth of *n*.

measure for the incomplete data produces a graph in Figure 4 that shows Kendall demonstrating a higher rate of convergence compared to the absolute distance measure. This phenomenon is caused by the growth of variability size of Kendall in incomplete data that grows exponentially with the growth of n, a similar condition to Spearman for complete data, while the variability size of the absolute distance measure is the same with the complete condition as shown by Figure 5. In this case, the absolute distance measure presents a ripple for small n, i.e. $n \leq 6$. Meanwhile, Kendall still presents a significantly different characteristics to the absolute distance measure for large n with $\hat{\sigma}_K^2 > \hat{\sigma}_c^2$ due to a wider range of Kendal's score in $[-1, 1]$ compared to the range of the absolute distance measure in $[0, 1]$. This condition is disclosed by graphing the variability function of Kendall for incomplete

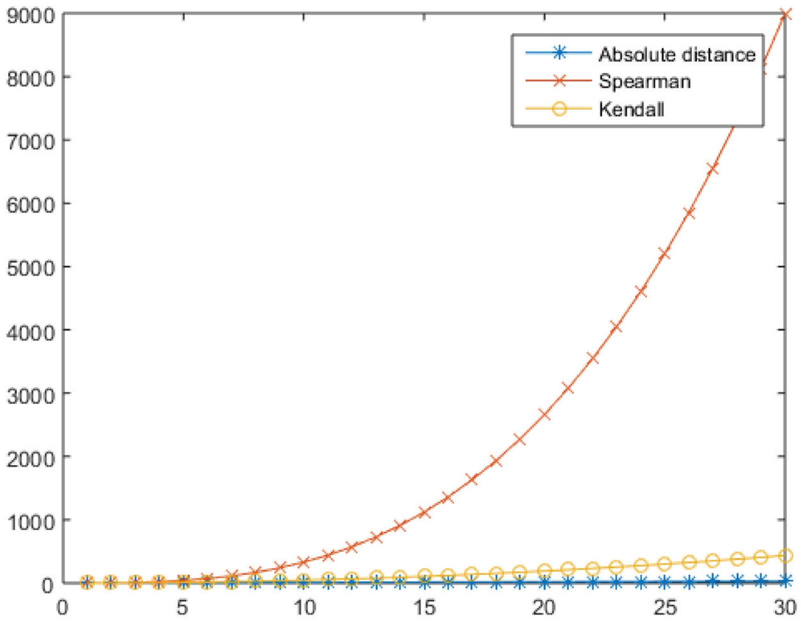
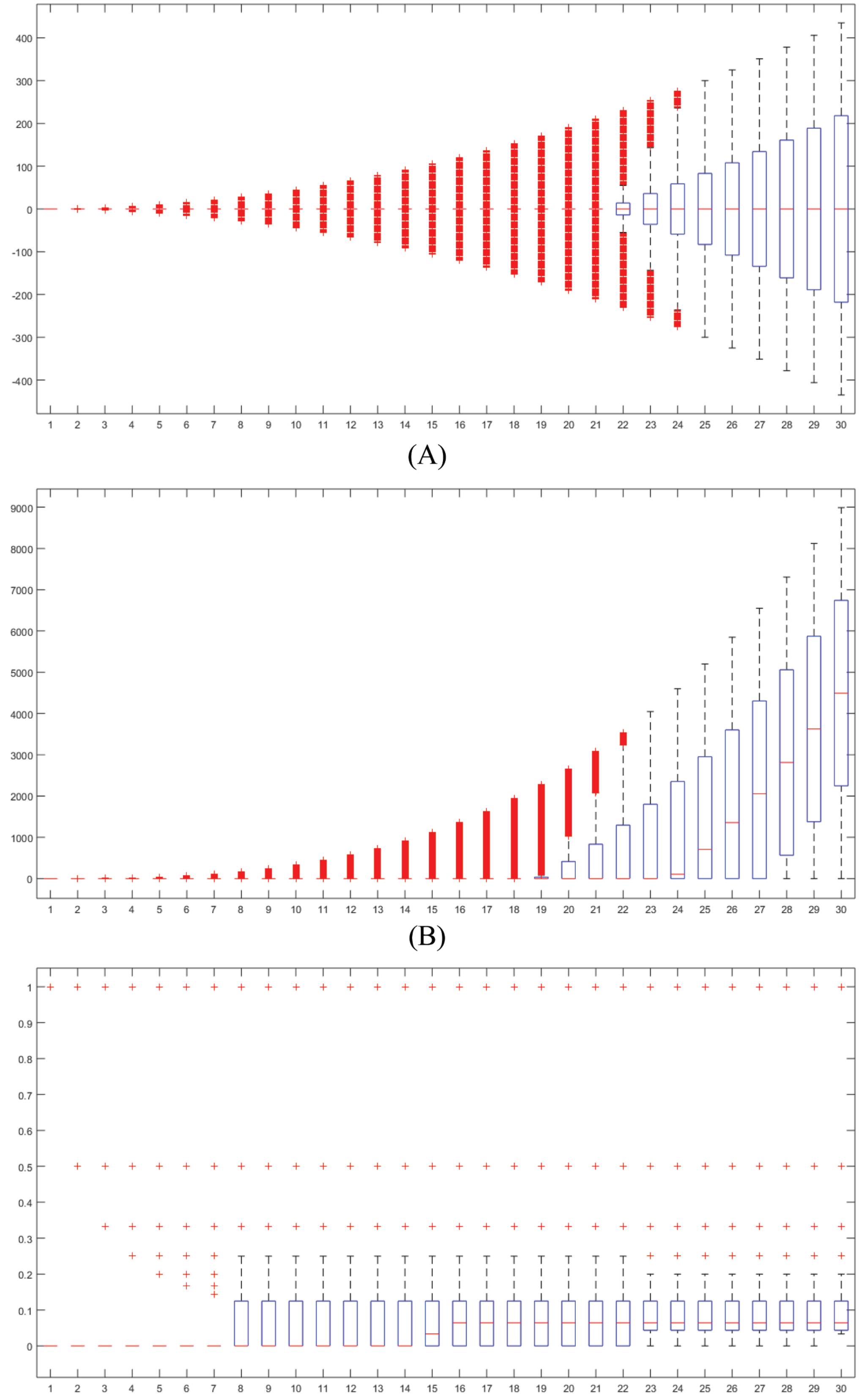


Figure 2. The variability size of kendall, Spearman, and the absolute distance measure based on the growth of *n* for complete data.



(A)

(B)

(C)

Figure 3. The variability distribution for (A) Kendall, (B) Spearman, and (C) The absolute distance measure.

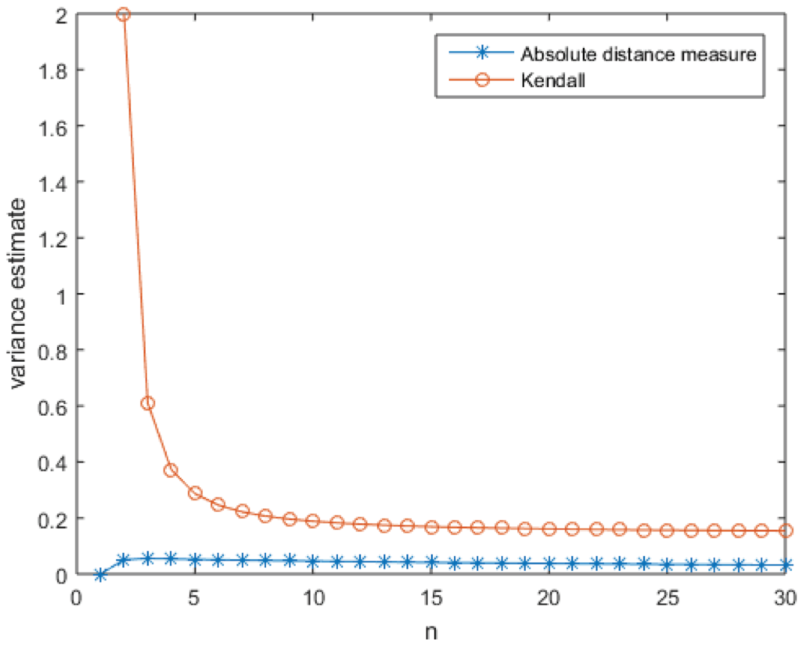


Figure 4. Variance estimates for Kendal and the absolute distance measure under incomplete condition based on the growth of *n*.

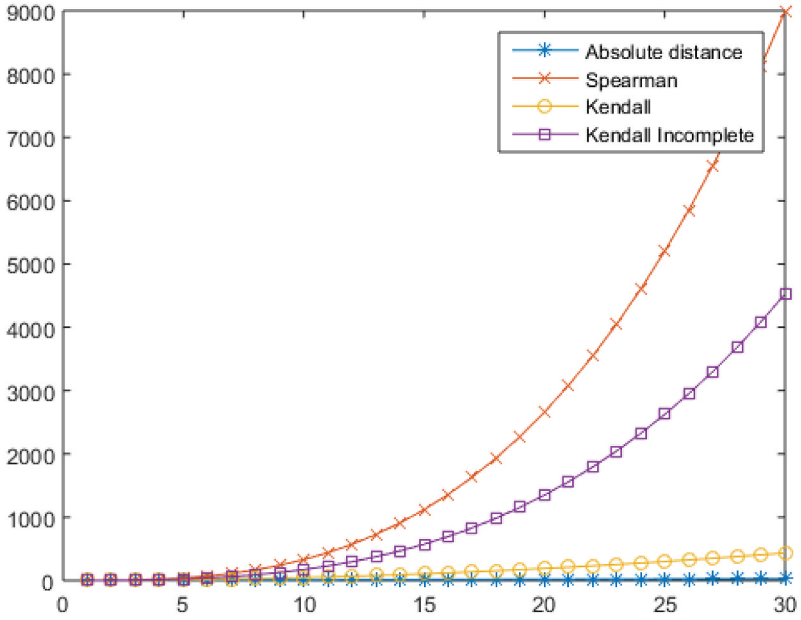


Figure 5. The exponential growth of Kendall's variability size for incomplete data compared to the original Kendal and Spearman in complete data and the absolute-distance measure.

data as shown by Figure 6 that shows Kendall still preserving the distribution of variability function similar to the complete data, even though the mean is moved away from zero by the growth of *n* due the incomplete condition. While the variability function of the absolute distance measure for the incomplete data is the same with the complete data as presented in Figure 3C that occupies the range $[0,1]$.

Meanwhile, the comparisons between the variance obtained from the complete data versus the incomplete condition for both Kendall and the absolute distance measure as given in Figure 7A and 7B, respectively, disclose a fact that the incomplete condition produces smaller variance than its counterpart. This condition is due to the larger number of possible data sequences for incomplete data as shown by Table 1 occupying the same range

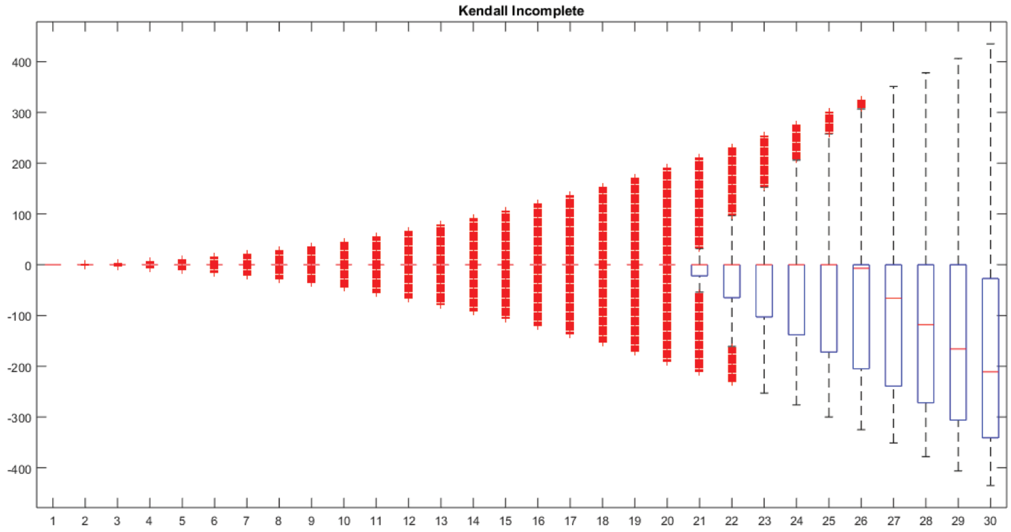
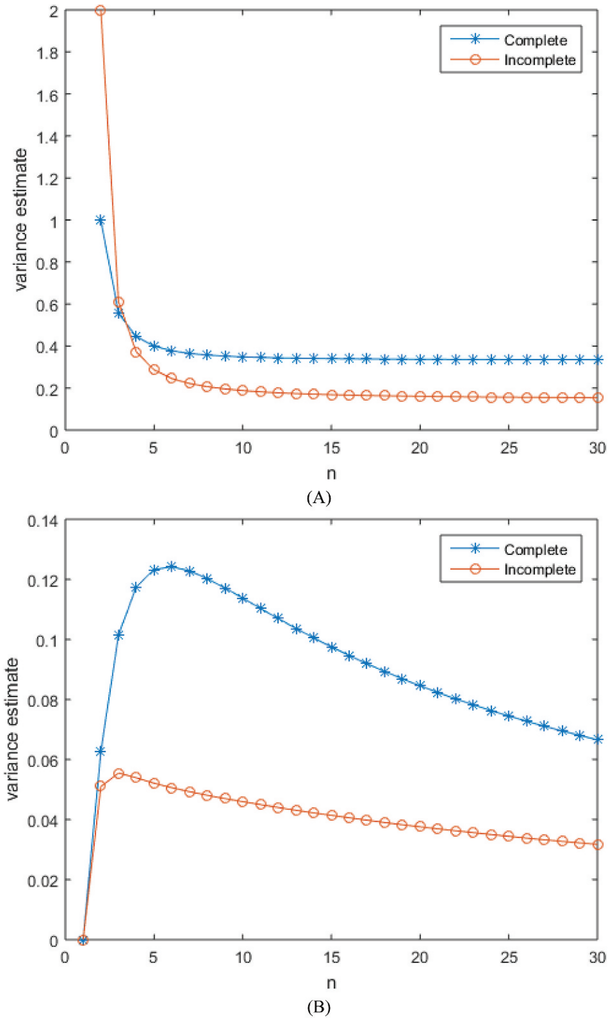


Figure 6. The variability function of Kendall for incomplete data.



(A)

(B)

Figure 7. Variance estimates for complete vs incomplete condition for (A) Kendall and (B) The absolute distance measure.

of variability as the complete data. For the case of Kendal, the complete data demonstrate higher rate of convergence than the incomplete condition as shown in Figure 7A. This phenomenon can be explained by contrasting the variability function of Kendall in both complete and incomplete condition as shown in Figure 3A and 6. In this case, even though the distribution of variability function in both condition is similar, however, the growth of data sequence in incomplete condition as described by Section 2A forces the mean to move away from zero by the growth of n. This condition causes the extension to achieve the convergence by Kendall in incomplete data.

While for the case of the absolute distance measure, both the complete and incomplete data show a similar rate of convergence as shown by Figure 7B. This condition is due to the same distribution of variability function as shown by Figure 3C and the same variability size as shown by Figure 2 and 5 for complete and incomplete data, respectively. In this case, the lower slope of variance estimate in incomplete data as shown by Figure 7B is due to the larger size of incomplete data sequences compared to the complete condition as described in Section 2A.

Conclusion

The variability function of rank correlation coefficients enables the expression of the mean and variance under the complete and incomplete data. It helps to solve the mathematical tractability of variance as noted by Xu et al. (2010). Here, we find that Kendall becomes the better method over Spearman and the absolute distance measure due to the following reason. Under the complete data condition, Kendall presents the variability function that perfectly distributes around the zero mean, which is different from other methods that fail to preserve the zero mean. In this case, Spearman shows the distribution of variability function above zero, while the absolute distant measure is in $[0,1]$. Although Spearman gains a higher convergence rate of variance, the distribution of variability function proves that Kendall is a well-designed method to conduct rank correlation for complete data. Under the incomplete condition, Kendall and the absolute distant measure exhibits the capacity to adapt with the missing data. In this case, Kendall needs to improve the definition of concordant and discordant by accepting alpha in order to deal with the missing data, while the absolute distant measure has this mechanism in its original design. However, Kendall shows a higher convergence rate than the absolute distant measure under the incomplete condition, and is able to preserve the zero mean of variability distribution to some extent of n. Meanwhile, Spearman fails to deal with the incomplete data. A lesson learned from Spearman incapability to deal with the missing data is due to the metrics that relies only on a direct measurement without preparing to deal with the unmeasured condition. Therefore, it is crucial to develop a normalization approach inside the metrics in order to adapt with varying input that potentially range from the normal to the extreme or beyond the normal condition. On the other hand, Kendall's indirect measurement to compute each data position in a ranking through the number of concordant and discordant survives the coefficient from the incomplete condition.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author received no direct funding for this research.

Notes on contributor

Cahyo Crysdian obtained bachelor degree in Electrical Engineering from Brawijaya University, Indonesia in 1997, and graduated from both master study and PhD in Computer Science from Universiti Teknologi Malaysia in 2003 and 2006 respectively. Currently, he works as a Head of Master Program in Computer Science, Islamic State University Maulana Malik Ibrahim Malang, Indonesia. His research interest includes Computer Vision and Intelligent System.

ORCID

Cahyo Crysdian http://orcid.org/0000-0002-7488-6217

References

Abdel-Aty, A.-H., Kadry, H., Zidan, M., Zanaty, E. A., Abdel-Aty, M., & Abdel-Aty, M. (2020). A quantum classification algorithm for classification incomplete patterns based on entanglement measure. *Journal of Intelligent and Fuzzy Systems*, *38*(3), 2817–2822. https://doi.org/10.3233/JIFS-179566

Albers, W., & Teulings, M. F. (1996). A simple estimator for correlation in incomplete data. *Statistics & Probability Letters*, *31*(1), 51–57. https://doi.org/10.1016/S0167-7152(96)00013-2

Alvo, M., & Cabilio, P. (1995). Rank correlation methods for missing data. *The Canadian Journal of Statistics, La Revue Canadienne de Statistique*, *23*(4), 345–358. https://doi.org/10.2307/3315379

Alvo, M., & Park, J. (2002). Multivariate non-parametric tests of trend when the data are incomplete. *Statistics &*

Probability Letters, *57*(3), 281–290. https://doi.org/10.1016/S0167-7152(02)00062-7

Cabilio, P., & Tilley, J. (1999). Power calculations for tests of trend with missing observations. *Environmetrics*, *10*(6), 803–816. https://doi.org/10.1002/(SICI)1099-095X(199911/12)10:6<803::AID-ENV397>3.0.CO;2-O

Crysdian, C. (2018). Performance measurement without ground truth to achieve optimal edge. *International Journal of Image and Data Fusion*, *9*(2), 170–193. https://doi.org/10.1080/19479832.2017.1384764

Eekhout, I., Enders, C. K., Twisk, J. W. R., de Boer, M. R., de Vet, H. C. W., & Heymans, M. W. (2015). Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 588–602. https://doi.org/10.1080/10705511.2014.937670

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(140), 1–37. https://doi.org/10.1186/s40537-021-00516-9

Kendall, M. G. (1938, June). A new measure of rank correlation. *Biometrika*, *30*(1–2), 81–93. https://doi.org/10.1093/biomet/30.1–2.81

Kidwell, P., Lebanon, G., & Cleveland, W. S. (2008, December). Visualizing incomplete and partially ranked data. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1356–1363. https://doi.org/10.1109/TVCG.2008.181

Kim, J., & Im, J. (2018). Proposing a missing data method for hospitality research on online customer reviews: An application of imputation approach. *International Journal of Contemporary Hospitality Management*, *30*(11), 3250–3267. https://doi.org/10.1108/IJCHM-10-2017-0708

Kim, J., Tae, D., & Seok, J., "A survey of missing data imputation using generative adversarial networks," *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2020, pp. 454–456, https://doi.org/10.1109/ICAIIC48513.2020.9065044.

Mirzaei, A., Carter, S. R., Patanwala, A. E., & Schneider, C. R. (2022). Missing data in surveys: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *18*(2), 2308–2316. https://doi.org/10.1016/j.sapharm.2021.03.009 1551-7411

Raykov, T., Schneider, B. C., Marcoulides, G. A., & Lichtenberg, P. A. (2014). Examining measure correlations with incomplete data sets. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 318–324. https://doi.org/10.1080/10705511.2014.882696

Rejeb, S., Duveau, C., & Rebafka, T. *"Self-organizing maps for exploration of partially observed data and imputation of missing values"*. Cornell University. arXiv:2202.07963, Feb 2022 https://doi.org/10.48550/arXiv.2202.07963

Spearman, C. (1904, January). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159

Szmidt, E., & Kacprzyk, J. (2011). *"The spearman and kendall rank correlation coefficients between intuitionistic fuzzy sets"*. EUSFLAT-LFA 2011 Atlantis Press EUSFLAT-LFA 2011

Xu, W., Hung, Y. S., Niranjan, M., & Shen, M. (2010, February). Asymptotic mean and variance of gini correlation for bivariate normal samples. *IEEE Transactions on Signal Processing*, *58*(2), 522–534. https://doi.org/10.1109/TSP.2009.2032448

Yan, A., Wang, W., Ren, Y., & Geng, H. W. (2021, June). A clustering algorithm for multi-modal heterogeneous big data with abnormal data. *Frontiers in Neurorobotics*, *15*, 1–9. https://doi.org/10.3389/fnbot.2021.680613

Zidan, M., Abdel-Aty, A.-H., El-Sadek, A., Zanaty, E. A., & Abdel-Aty, M., "Low-cost autonomous perceptron neural network inspired by quantum computation", *AIP conference proceedings*, 2017, 1905, 020005, https://doi.org/10.1063/1.5012145.