Representative Model the Graph Theory in Calculations
Kendall Correlation Coefficient

1Pasukat Sembiring, 2Ujian Sinulingga, 3Marihat Situmorang, 4Sajadin Sembiring.
1,2,3.Department of Mathematics, University of Sumatera Utara
4Department of Computer Science, University of Sumatera Utara

Abstract. Kendall correlation coefficient formula as an association or relationship size measure between two ordinal scale data variables, has been widely used in research in various fields of science. A study of the Kendall correlation coefficient formula associated with the graph theory through a complete asymmetric graph as an adjacency matrix, will be used in Kendall correlation calculations in the case of bivariate. The results of this study indicate that a complete asymmetric graph as an adjacency matrix can be used as an alternative method of Kendall correlation coefficient calculation.

Keyword: Graph Theory, Kendall Correlation Coefficient, adjacency matrix.

1. Introduction
Correlation is a study that discusses the degree of relationship between two or more variables [1]. Correlation is one of the statistical analysis techniques that are widely used by researchers because researchers are generally interested in the events that occur and try to connected. The degree of closeness of the relationship between two or more variables can be known by looking for the amount of correlation numbers commonly called the coefficient correlation.

Nonparametric statistical methods are often called distribution free methods because their statistical test models do not specify certain conditions about the distribution of population parameters [2]. The nonparametric statistics test specifies only the assumption that the observations should be independent and that the variables under study must essentially have continuity. Many of the nonparametric statistical tests are sometimes referred to as rank tests, since these nonparametric techniques can be used for scores that are not exact scores in the sense of scarcity, but rather scores that are solely ranks.

One method of measuring nonparametric correlation coefficient is the rank correlation coefficient of Kendall. The rank correlation coefficient of Kendall was first proposed by Maurice G. Kendall in 1938. Coefficient of rank correlation Kendall is denoted by τ. Kendall rank correlation coefficient (τ) is used to find the relationship and test the hypothesis between two or more variables, if the data is ordinal or rank.
The degree of closeness between the two variables can be shown by the ratio (comparison) between the actual score of +1 and -1 with the maximum score that can be achieved. +1 is assigned to a naturally arranged pair and -1 is assigned to a pair that is not naturally composed.

The rank correlation coefficient of Kendall (τ) has advantages when compared with spearman rank correlation coefficient (τs). τ is more general because it can be calculated as normal distribution and can be searched for the coefficient correlation [3].

Graph theory was first introduced by Leonard Euler in 1736 when he proved the possibility of passing through four areas linked to seven bridges over the Pregel river in Konigsberg, Russia in one pass through each bridge just once and returning to its original place [4]. The Konigsberg bridge problem can be expressed in graph by determining the four regions as vertices and the seven bridges as the edges connecting the corresponding vertex pairs.

The calculation τ using graph theory is done by forming complete asymmetric digraph with vertex is every object of research [5]. Complete asymmetric digraph and then poured into adjacency matrix. From the adjacency matrix formed can be calculated actually score. And for the maximum score that can be achieved is equal to the number of edges with n vertices on the complete asymmetric digraph. The number of arcs can be obtained from \( C_n^2 \). So that the correlation coefficient rank of Kendall (τ) is the ratio actual score with maximum score that can be achieved.

2. Methodology

The steps to determine the correlation coefficient of rank Kendall (τ) through graph theory approach that is:
1. Establishing a complete asymmetric digraph with vertices is every object of the research.
2. Establish adjacency matrix obtained from complete asymmetric digraph.
3. Calculating actual score of adjacency matrix formed.
4. Calculate the maximum possible score. Score maximum is equal to the number of arcs with n vertex on complete asymmetric digraph. The number of arcs can be obtained from \( C_n^2 \).
5. Subordinate actual score and maximum possible score into the correlation rank correlation coefficient formula (τ) with approach through graph theory.

2.1. Correlation Coefficient Rank Correlation (T)

The rank correlation coefficient of Kendall (τ) can be obtained by comparing actual score with maximum score possible [6]. Or in other words the actual score is the actual score +1 and -1. +1 is assigned to a naturally arranged pair and -1 is assigned to a pair that is not naturally composed. Whereas the maximum possible score is determined by an arrangement \( C_n^2 \) that can be decomposed into \( \frac{1}{2} n(n-1) \). So, Kendall rank correlation coefficient (τ) can be formulated:

\[
\tau = \frac{\text{actual score}}{\text{maximum score}}
\]

Furthermore the actual score is given the symbol S, and the maximum score is determined by the arrangement \( C_n^2 \), where n is the number of objects or individuals in the random variables X and Y. Can be mathematically written:

\[
\tau = \frac{S}{C_n^2}
\]

Or
\[ \tau = \frac{S}{2n(n-1)} \]

Where  
\( S \) = actual score (score count +1 and -1)  
\( N \) = number of objects or individuals on random variables X and Y

There are times when random variables X and Y have the same object or often called the twin rank. If there are two or more observed values (either for the same random variable X or Y), then those values are given average rank. The effect of the twin rank values is to change the value of the denominator on the rank correlation coefficient formula of Kendall (\( \tau \))[7]. In this case the correlation rank correlation coefficient formula (\( \tau \)) becomes:

\[ \tau = \frac{S}{\sqrt{\frac{1}{2}n(n-1) - T_x}\sqrt{\frac{1}{2}n(n-1) - T_y}} \]

Where  
\( S \) = actual score (score count +1 and -1)  
\( N \) = number of objects or individuals on random variables X and Y  
\( T_x = \frac{1}{2} \sum t(t - 1) \)
\( T_y = \frac{1}{2} \sum t(t - 1) \)

\( T_x \) is the number of twin rank of each twin group for the variable X  
\( T_y \) is the number of twin rank for each of its twin groups for variable Y

2.2. Complete Asymmetric Digraph
Complete asymmetric digraph is an asymmetric digraph where there is exactly one between each pair of vertices. Complete asymmetric digraph with \( n \) vertices containing \( \frac{1}{2}N(N - 1) \) arc [8].

For evidence: Let D be a complete asymmetric digraph with \( n \) vertices \( v_1, v_2, ..., v_n \). Take any vertex (call \( v_1 \)). Since D is a complete asymmetric digraph, then \( v_1 \) is associated with \( (n - 1) \) other vertices (\( v_2, v_3, ..., v_n \)). So, there is \( (n - 1) \) arc fruit.

Next, grab any second vertex (call \( v_2 \)). Since D is complete asymmetric digraph, then \( v_2 \) is also associated with all remaining vertices (\( v_3, ..., v_n \)) so that there are \( (n - 2) \) arcs associated with \( v_2 \). The process is continued by counting the number of arcs associated with \( v_3, v_4, ..., v_{n-1} \) and which have not been calculated before. Many arcs obtained are: \( (n - 3), (n - 4), ..., 3,2,1 \).

So overall there is \( (n - 1) + (n - 2) + (n - 3) + ... + 3 + 2 + 1 = \frac{n(n - 1)}{2} \) arc

In the Figure 1 can be seen an example of complete asymmetric digraph.
Complete asymmetric digraph can be represented in an adjacency matrix. The number of adjacency matrix rows and columns is the same as the vertex in the complete asymmetric digraph. The corresponding adjacency matrix is the matrix of square n x n, ie the matrix $A = (a_{ij})$ by:

$$
(a_{ij}) = \begin{cases} 
+1 & \text{if any arc from } v_i \text{ to } v_j \text{ natural} \\
-1 & \text{if any arc from } v_i \text{ to } v_j \text{ unnatural} \\
0 & \text{if not a arc } v_i \text{ to } v_j 
\end{cases}
$$

3. Result and Discussion

3.1. Graph Form as An Adjacency Matrix in Determination of Rank Correlation Coefficient Kendall ($\tau$)

The graph used in the determination of the rank correlation coefficient of Kendall ($\tau$) is directed graph (digraph), because each object of observation in the given data is sequential paired data. Vertices are the objects in the study. In the Kendall rank correlation method, the observed objects are linked to each of the next objects from the first to the second, then from the first to the third, and so on until object n. After that proceed from second object to third object, second object to fourth object, and so on until object n. This continues through next n object to next n object. Or in other words, if the objects in the study is a vertex in the graph then from the above understanding can be written: $(v_i, v_j)$ with $i < j$; $i, j = 1, 2, 3, ..., n$.

$v_i$ is the starting point and $v_j$ is the end point, so an arc with the vertex $(v_i, v_j)$ represents the line from point $i$ and point $j$. Because $(v_i, v_j)$ is connected only once with $i < j$; $i, j = 1, 2, 3, ..., n$ so that there is no parallel line and no arc with end points $(v_i, v_j)$ then the graph that is set is an asymmetric digraph. And because each pair of objects in the study must be connected or in other words there is exactly one arc between each pair of vertices $(v_i, v_j)$ with $i < j$; $i, j = 1, 2, 3, ..., n$ then the graph form used in the T calculation is a complete asymmetric digraph.

Complete asymmetric digraph formed then poured into adjacency matrix. The number of bans and adjacency matrix columns is equal to the number of vertices in the complete asymmetric digraph. The corresponding adjacency matrix is the matrix of square n x n, in matrix $A = (a_{ij})$ by:

$$
(a_{ij}) = \begin{cases} 
+1 & \text{if any arc from } v_i \text{ to } v_j \text{ natural} \\
-1 & \text{if any arc from } v_i \text{ to } v_j \text{ unnatural} \\
0 & \text{if not have arc titik } v_i \text{ to } v_j 
\end{cases}
$$

So that the adjacency matrix is formed into:

\[
A = (a_{ij}) = \begin{bmatrix}
0 & a_{12} & a_{13} & a_{14} & \cdots & a_{1n} \\
0 & 0 & a_{23} & a_{24} & \cdots & a_{2n} \\
0 & 0 & 0 & a_{3n} & \cdots & a_{3n} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \cdots & \cdots & 0 \\
0 & 0 & \cdots & \cdots & \cdots & 0
\end{bmatrix}
\]

After the adjacency matrix is formed with the above conditions, then the actual score is determined by summing the $+1$ and $-1$ values obtained from the adjacency matrix. Then determine the maximum possible score. The maximum score may be equal to the number of arcs of complete asymmetric digraph,
so the maximum score can be determined by $C^S$ where $n$ is the number of vertices of the complete asymmetric digraph.

Thus, the rank correlation coefficient of Kendall ($\tau$) with the approach through graph theory can be determined by the formula:

$$\tau = \frac{S}{C^S} \quad \text{or} \quad \tau = \frac{S}{\frac{1}{2} n(n-1)}$$

For the rank Kendall correlation coefficient with twin rank then the formula used is:

$$\tau = \frac{S^2}{\sqrt{C^x - T_x} \sqrt{C^y - T_y}}$$

Where, $S = $ actual score

$$T_x = \frac{1}{2} \sum t(t-1)$$

$T$ is the number of twin rankings of each twin group for the variable $x$

$$T_y = \frac{1}{2} \sum t(t-1)$$

$T$ is the number of twin rankings of each twin group for the $y$ variables

Furthermore, we can develop null hypothesis and alternative hypothesis that can be tested by hypothesis test using formula:

$$Z = \frac{\tau}{\sigma_t}$$

Where:
- $Z = $ statistical test for the correlation rank
- $\tau = $ the rank correlation coefficient of Kendall
- $\sigma_t = $ standard deviation, can be obtained by formulation:

$$\sigma_t = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$$

3.2. Model Usage

As a test of Kendall rank correlation coefficient calculation model ($\tau$) using graph theory, exemplified by the following data which is only a simulation data. The simulation data was taken from the second edition of nonparametric statistical book by Samubara Saleh. The following is the observation data of the chief physician of the internal medicine section of 20 patients with high blood pressure.

| No. | Patient | Age | Tention |
|-----|---------|-----|---------|
| 1   | A       | 56  | 147     |
| 2   | B       | 42  | 125     |
| 3   | C       | 72  | 160     |
| 4   | D       | 36  | 118     |
| 5   | E       | 63  | 149     |
| 6   | F       | 47  | 128     |
| 7   | G       | 55  | 150     |
| 8   | H       | 49  | 145     |
To get the maximum score from the above data can be obtained from the maximum arc amount obtained from complete asymmetric digraph. 20 patients with high blood become vertex of complete asymmetric digraph. The maximum score is equal to the number of arcs of complete asymmetric digraph, so the maximum score can be determined by \( C_n \) where \( n \) is the number of vertices of the complete asymmetric digraph. From the data, it is known that the number of \( n \) is 20, so the maximum score can be obtained from the number of arcs of the complete asymmetric digraph using theorem 2, 4, 1, then the maximum score can be obtained by:

\[
C_{20}^2 = \frac{1}{2} 20(20 - 1) = \frac{380}{2} = 190
\]

Can be presented with digraph will be obtained graph with 20 vertexes and 190 arcs. And can be seen in the following figure:

![Figure 2. Graph G (20, 190)](image)

So, the maximum score obtained is equal to the number of arcs of the complete asymmetric digraph of 190. To determine the actual score value can be obtained by using adjacency matrix in the form of upper triangular matrix.

**Table 2.** Ranking of Physician Observation of Internal Disease Section on 20 Patients with High Blood

| No. | Patient | Age | Tention |
|-----|---------|-----|---------|
| 1   | A       | 8   | 10      |
Next rank is arranged in order. The results are presented in the following table:

**Table 3. Ranking Arrangements Observation Doctors Head of Internal Disease Against 20 Patients Suffering from High Blood**

| No. | Pasien | Umur | Tekanan Darah |
|-----|--------|------|--------------|
| 1   | R      | 1    | 1            |
| 2   | C      | 2    | 4            |
| 3   | P      | 3    | 2            |
| 4   | K      | 4    | 6            |
| 5   | Q      | 5    | 3            |
| 6   | E      | 6    | 8            |
| 7   | L      | 7    | 5            |
| 8   | A      | 8    | 10           |
| 9   | G      | 9    | 7            |
| 10  | M      | 10   | 9            |
| 11  | H      | 11   | 11           |
| 12  | F      | 12   | 14           |
| 13  | S      | 13   | 13           |
| 14  | N      | 14   | 15           |
| 15  | B      | 15,5 | 16,5         |
| 16  | J      | 15,5 | 12           |
| 17  | Q      | 5    | 3            |
| 18  | R      | 1    | 1            |
| 19  | S      | 13   | 13           |
| 20  | T      | 20   | 19,5         |

After the age rank is sorted naturally, the next step is to establish an adjacency matrix corresponding to the complete asymmetric digraph.
From adjacency matrix above calculated total value of +1 and -1. From here obtained actual score. The actual score can be seen from the table below. Each row in this table shows the number of +1 values (naturally arranged pair) and the number of values -1 (naturally incomplete pair) of each row in the adjacency matrix.

**Table 4.** Couples participating naturally (+1) and unbranded pair (-1) of each line in adjacency matrix

| Many Natural Ordered Pairs (+1) | Many Unnoticed Pairs Naturally (-1) |
|---------------------------------|-------------------------------------|
| 19                              | 0                                   |
| 16                              | -2                                  |
| 17                              | 0                                   |
| 14                              | -2                                  |
| 15                              | 0                                   |
| 12                              | -2                                  |
| 13                              | 0                                   |
| 10                              | -2                                  |
| 11                              | 0                                   |
| 10                              | 0                                   |
| 9                               | 0                                   |
| 6                               | -2                                  |
| 5                               | -1                                  |
| 3                               | -1                                  |
| 4                               | 0                                   |
| 3                               | 0                                   |
| 0                               | -1                                  |
| 1                               | 0                                   |

174    14

From the above table the actual score obtained by adding a lot of couples are sequenced naturally with many couples are not arranged naturally.

\[ S = 174 - 14 \]

\[ S = 160 \]
Because that will be tested has twin rank, then the next search result of Tx and Ty. Where the observation of age is given by x and the observation of blood pressure is given by y. The rank of observation of age there is one group of twin values that is: rank 15.5 with t = 2. So, the value of Tx is:

$$T_x = \frac{1}{2} \sum t(t - 1) = \frac{1}{2} [2(2 - 1)] = \frac{1}{2} [2] = 1$$

In the observation rank of blood pressure there are two groups of twin values: rank 16.5 with 1 = 2 and rank 19.5 with t = 2. So, the value of Ty is:

$$T_y = \frac{1}{2} \sum t(t - 1) = \frac{1}{2} [2(2 - 1) + 2(2 - 1)] = \frac{1}{2} [2 + 2] = 2$$

Thus, the Kendall rank correlation coefficient value can be searched using the formula:

$$\tau = \frac{S}{\sqrt{C_x^2 - T_x} \sqrt{C_y^2 - T_y}} = \frac{160}{\sqrt{190 - 1} \sqrt{190 - 2}} = \frac{160}{\sqrt{188}} = \frac{160}{13,747} \approx 0.118$$

Further testing against \(\tau\):

1. Determine the null hypothesis and alternative hypothesis, yaitu :
   - \(H_0\): there is no correlation between age and blood pressure
   - \(H_1\): there is a correlation between age and blood pressure

2. Criterion of decision making:
   - \(H_0\) : accepted and \(H_1\) is rejected if \(-Z_{\alpha/2} \leq Z_H \leq Z_{\alpha/2}\)
   - \(H_0\) : rejected and \(H_1\) is rejected if \(Z_H > Z_{\alpha/2}\) or \(Z_H < -Z_{\alpha/2}\)

3. The value of Z arithmetic is:

$$Z = \frac{\tau}{\sqrt{9(n-1)} = \frac{0.8488}{\sqrt{9(20)(20-1)}} = \frac{0.8488}{\sqrt{3420}} = 0.0263 \approx 0.1622$$

4. The critical value at \(\alpha = 5\% = \pm Z_{1/2(1-\alpha)} = Z_{1/2(1-0.05)} = Z_{0.975} = \pm 1.96\) (use normal curve)

5. Contribution
   - \(H_0\) can not to process because \(Z_H = 5,2330 > Z_{0.025} = 1.96\)
   - Meaning: there is a significant correlation between age with blood pressure from 20 patients with high blood pressure.

4. Conclusion
   To determine the maximum score can be obtained from the accumulation of aii arc of the complete asymmetric digraph and to obtain actual score can be obtained by using the adjacency matrix of complete asymmetric digraph in the form of upper triangular matrix. So, from the results of the above discussion can be concluded that the graph directed can be used to determine the value of correlation coefficient rank Kendall (\(\tau\)).

Reference
[1]. Frey, L., Botan, C., & Kreps, G. (1999). Investigating communication: An introduction to research methods. (2nd ed.) Boston: Allyn & Bacon.
[2]. Siagian. 2000. Metode statistika untuk bisnis dan ekonomi. Jakarta: PT Gramedia Pustaka Utama
[3]. Djarwanto Ps. Edisi 2004. Statistika Nonparametrik. Yogyakarta: BPFE-Yogyakarta
[4]. Alavi, Y. 1985. *Graph Theory -with Applications to Algorithms and Computer Science*. USA: John Wiley & Sons, Inc

[5]. Beineke, Lowell W & Wilson, Robin J. 1983. *Selected Topics in Graph Theory* 2. New York: Academic Press Inc.

[6]. Wilson, Robin J. 1985. *Introduction to Graph Theory*. Third Edition. New York: Longman, Inc

[7]. Narsingh, Deo. 1984. *Graph Theory with application to Engineering and Computer Science*. New Delhi: Prentice Hall of India, Private Limited.

[8]. Narsingh, Deo. 1984. *Graph Theory with application to Engineering and Computer Science*. New Delhi: Prentice Hall of India, Private Limited.

[9]. Saleh, Samsubar. Drs. M.Soc. 1996. *Satistik Nonparametrik Edisi 2*. Yogyakarta: Universitas Gajah Mada