A Modulo-Based Computation for Kendall’s Phi Test for 2 by 2 Contingency Table

Ray-Ming Chen
School of Mathematics and Statistics, Baise University, 21, Zhongshan No.2 Road, Guangxi Province, China
E-mail: baotaoxi@163.com

Abstract. Kendall’s Phi Test is very important technique in testing whether there exists a significant relation between variables under some known levels of experiments. However, the typical statistical testing would utilise some complex probability distribution for the statistic. In this article, we put forward a computational approach to effectively linearize Kendall’s Phi coefficients based on modulo operation. In our approach, the confidence levels would serve the threshold of the operation. Unlike typical approach which lists the whole distribution of the statistic, our method would concentrate only on the accumulated values needed. Such computation has merits in visualization and programming. In particular, if one is interested in Kendall’s Phi-based test, then this approach could largely reduce the complexity and computational time.

1. Introduction
Kendall’s Phi statistic [1, 2] plays an important role is statistical hypothesis testing [3, 4] for contingency table, which is a common form in data analysis. However, due to the complexity of this statistic, the computation would be way too complex for such computation. We adopt modulo operation to linearize the value of Kendall’s Phi statistic. Such linearization facilitates the computation of exact Kendall’s Phi statistic, rather than to estimation based on normal distribution [5, 6].

In this article, we aim to largely reduce such complexity and give an effective approach to tackle such computation. Unless otherwise specified, all the variables in this article denotes non-negative integers. A contingency table is in the form:

|       | Quality One | Quality Two | Sum |
|-------|-------------|-------------|-----|
| Experiment One | $x_{11}$ | $x_{12}$ | $r_1$ |
| Experiment Two  | $x_{21}$ | $x_{22}$ | $r_2$ |
| Sum            | $c_1$    | $c_2$    | $N$  |

The above-mentioned variables respect the following equations: $x_{11} + x_{12} = r_1$, $x_{21} + x_{22} = r_2$, $x_{11} + x_{21} = c_1$, $x_{12} + x_{22} = c_2$, $r_1 + r_2 = c_1 + c_2 = N$, where $r_1$, $r_2$ and $N$ are all predetermined.
kendall’s Phi statistic is defined as

$$\phi = \sqrt{\frac{N}{r_1 \cdot r_2} \cdot \left( \frac{x_{11} \cdot x_{22} - x_{12} \cdot x_{21}}{c_1 \cdot c_2} \right)}.$$  

(1)

Observe that $0 \leq x_{11} \leq \min\{r_1, c_1\}$ and $0 \leq x_{22} \leq \min\{r_2, c_2\}$. Later on, we will apply these properties in our derivations.

2. Modulo Computation and Claims

Let $\mathbb{N}$ denote the set of natural numbers. Let $cd$ denote the multiplication $c \cdot d$. Let $A = \{a_1, a_2, \ldots, a_{|A|}\}, B = \{b_1, b_2, \ldots, b_{|B|}\} \subseteq \mathbb{N}$ be an arbitrary, where $|A|$ denotes the size of set $A$. We use the notations $c \cdot A$ (or simply $cA$) and $c + A$ to denote the sets $\{c \cdot a_1, c \cdot a_2, \ldots, c \cdot a_{|A|}\}$ and $\{c + a_1, c + a_2, \ldots, c + a_{|A|}\}$, respectively. Moreover, $c \cdot A + d \cdot B$ (or $cA + dB$) denotes the set $\{c \cdot a + d \cdot b : a \in A, b \in B\}$. We use $\gcd(c, d)$ to denote the greatest common divisor of $c$ and $d$. Furthermore, $[m]$ is used to denote the set $\{0, 1, 2, \ldots, m\}$. Hence $n[m] = \{n \cdot 0, n \cdot 1, n \cdot 2, \ldots, n \cdot m\}$.

If (by Division Theorem) $c = dq + r$, we use $Q_d(c)$ to denote the quotient $q$ and $R_d(c)$, the remainder of $r$. If all $a \in A, b \in B[a \leq b]$, it is denoted by $A \leq B$. For any non-negative real number $\gamma$, we use $[\gamma]$ and $[\gamma]$ to denote the greatest integer less than or equal to $\gamma$ and the smallest integer greater than or equal to $\gamma$, respectively.

Claim 1. Given natural numbers $1 \leq m \leq n$ and $\gcd(m, n) = 1$, then for all $c, d \leq m[c \neq d \iff c \cdot m \mod n \neq d \cdot m \mod n]$.

Claim 2. Given natural numbers $1 \leq m \leq n$ and $\gcd(m, n) = 1$, then

$$n[n] + m[m] = \bigcup_{k=0}^{n} \{(k + Q_n(m \cdot i)) \cdot n + R_n(m \cdot i) : 0 \leq i \leq m\}.$$

Claim 3. $nk + mi \mod n = nk' + mi' \mod n$ iff $k = k', i = i'$.

Definition 2.1. Define a partial inequality $\leq$ on $\mathbb{N} \times \mathbb{N}$ by $(n_1, n_2) \leq (m_1, m_2)$ iff $n_1 < m_1$ or $[n_1 = m_1, n_2 \leq m_2]$. Furthermore, define a strict partial inequality $<$ on $\mathbb{N} \times \mathbb{N}$ by $(n_1, n_2) < (m_1, m_2)$ iff $[n_1 = m_1, n_2 \leq m_2]$. Hence $(n_1, n_2) \leq (m_1, m_2), (n_1, n_2) \neq (m_1, m_2)]$.

Claim 4. $c \leq d$ iff $(Q_n(c), R_n(c)) \leq (Q_n(d), R_n(d))$.

Kendall’s Phi statistics could be represented by the following form:

Claim 5. $\phi = \sqrt{\frac{N}{r_1 \cdot r_2} \cdot \frac{(c_2 \cdot x_{11} + (N - c_2) \cdot x_{22} - (N - c_2) \cdot c_2}{(N - c_2)^2}}.$

Claim 6. $P\left(\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}\right) = p^N \cdot \left(\begin{bmatrix} r_1 \\ x_{11} \end{bmatrix}\right) \cdot \left(\begin{bmatrix} r_2 \\ x_{22} \end{bmatrix}\right) \cdot (\frac{1}{p} - 1)^{c_2}$, where $P$ denotes the joint probability of the contingency table and $p$ is some given probability.

Claim 7. If one interchanges $c_1$ and $c_2$, then he has the opposite signs of their corresponding resulting probabilities.

3. Algorithms

In order to linearize the values of Kendall’s Phi statistic and facilitate the computation of their corresponding probabilities, we specify several algorithms and theories in this section. We use $\vec{v}$ to denote a vector and $\vec{v}(n)$ to denote its $n$-th element. Each $\vec{v}$ would be in the following form: $\{x_{22}, Q_{N-c_2}(c_2 \cdot x_{11}); R_{N-c_2}(c_2 \cdot x_{11}), c_2, x_{11}\}$, where $0 \leq x_{11} \leq \min\{r_1, N - c_2\}$ and $0 \leq x_{22} \leq \min\{r_2, c_2\}$ to keep track of all the information needed for our computation, which could be envisaged via Claim 5 and 6. For better sighting purpose, we use one semi-comma
to denote a comma in the vector \( \vec{v} \). The whole mechanism could be executed based on the induction on \( c_2 \). Because of Claim 7, one only needs to take values of \( c_2 \) from 0 to \( \lceil \frac{N}{2} \rceil \). The algorithm goes as follows:

(i) Compute the set \( \Omega^{c_2} = \{(Q_{c_1}(c_2 \cdot x_{11}), R_{c_1}(c_2 \cdot x_{11})) : 0 \leq x_{11} \leq \min\{r_1, c_1\}\} \).

(ii) \( KS_{c_2}(0) = \{(0,0; R_{c_1}(c_2 \cdot x_{11}), c_2, x_{11}) : (0, R_{c_1}(c_2 \cdot x_{11})) \in \Omega^{c_2}\} \).

(iii) \( KS_{c_2}(1) = \{(1,1; 0,0,0) + [\vec{v} \in KS_{c_2}(0) : \vec{v}(1) \leq \min\{r_2, c_2\}] \cup \{(0,1; R_{c_1}(c_2 \cdot x_{11}), c_2, x_{11}) : (1, R_{c_1}(c_2 \cdot x_{11})) \in \Omega^{c_2}\} \).

(iv) Suppose \( KS_{c_2}(t) \) is defined, then one defines \( KS_{c_2}(t+1) := \)

\[
[(1,1; 0,0,0) + [\vec{v} \in KS_{c_2}(t) : \vec{v}(1) \leq \min\{c_2, r_2\}] \cup \{(0,t+1; R_{c_1}(c_2 \cdot x_{11}), c_2, x_{11}) : (t + 1, R_{c_1}(c_2 \cdot x_{11})) \in \Omega^{c_2}\},
\]

until \( t + 1 = 1 + \min\{c_2, r_2\} + \left\lfloor \frac{c_2 \min\{r_1, c_1\}}{c_1} \right\rfloor \).

Then we define a set \( KS_{c_2} \) to be the union of all the above sets, i.e.,

\[
KS_{c_2} = \bigcup_{n=1}^{\lceil \frac{N}{2} \rceil} KS_{c_2}(n).
\]

Let \( KS \) denotes \( \bigcup_{c_2=0}^{\lceil \frac{N}{2} \rceil} KS_{c_2} \).

**Example 1.** Suppose \( N = 12, r_1 = 4, c_2 = 4 \). Then \( c_1 = 8, r_2 = 8 \) and \( \Omega^4 = \{(0,0), (0,4), (1,0), (1,4), (2,0)\} \). The execution of the algorithm could be demonstrated in the following:

- \( KS_4(0) = \{(0,0; 0,4,0), (0,0; 4,1,4)\}\)
- \( KS_4(1) = \{(1,1; 0,4,0), (1,1; 4,1,4), (0,1; 0,4,2), (0,1; 4,4,3)\}\)
- \( KS_4(2) = \{(2,2; 0,4,0), (2,2; 4,4,1), (1,2; 0,4,2), (1,2; 4,4,3), (0,2; 0,4,4)\}\)
- \( KS_4(3) = \{(3,3; 0,4,0), (3,3; 4,4,1), (2,3; 0,4,2), (2,3; 4,4,3), (1,3; 0,4,4)\}\)
- \( KS_4(4) = \{(4,4; 0,4,0), (4,4; 4,4,1), (3,4; 0,4,2), (3,4; 4,4,3), (2,4; 0,4,4)\}\)
- \( KS_4(5) = \{(4,5; 0,4,2), (4,5; 4,4,3), (3,5; 0,4,4)\}\)
- \( KS_4(6) = \{(4,6; 0,4,4)\}\)

**Definition 3.1.** For each \( \vec{v} \in KS_{c_2}(n) \), define \( < \vec{v} > := \frac{c_2 \vec{v}(5) + c_1 \vec{v}(1) - c_1 \cdot c_2}{\sqrt{c_1 \cdot c_2}} \).

**Definition 3.2.** For each \( n \), we inductively define the following sets:

- \( K_{c_2}(n, 1) := \arg\min\{< \vec{v} > : \vec{v} \in KS_{c_2}(n)\}\)
- \( K_{c_2}(n, 2) := \arg\min\{< \vec{v} > : \vec{v} \in KS_{c_2}(n) - K_{c_2}(n, 1)\}\)
- ...  
- If the set \( K_{c_2}(n, j) \) is defined, then we define

\[
K_{c_2}(n, j + 1) := \arg\min\{< \vec{v} > : \vec{v} \in KS_{c_2}(n) - \bigcup_{i=1}^{j} K_{c_2}(n, i)\}.
\]

**Definition 3.3.** Let \( \mathbb{K}_j \) denote the vector whose elements are the ascending vectors in \( KS_{c_j} \).

We have to emphasize that each element in the vector \( \mathbb{K}_j \) is also a vector. This setting facilitates our instruction of the following minimal search method. In this method, we use \( Net^i \) to contain all the minimal values in each category \( KS_j \) and store the minimal ones in the set \( \tau^i \).
Definition 3.4. (Net-type Minimal Search)

- \(\text{Net}^1 := (0, 0, \ldots, 0); \tau^1 := \arg\min\{<\vec{v}>: \vec{v} \in \bigcup_{j=0}^{\infty} \vec{K}_j(\text{Net}^1(j))\}\)

- \(\text{Net}^2 = \text{Net}^1[\tau^1(4), \text{Net}^1(\tau^1(4)) + 1]; \tau^2 := \arg\min\{<\vec{v}>: \vec{v} \in \bigcup_{j=0}^{\infty} \vec{K}_j(\text{Net}^2(j))\}\)

- If \(\text{Net}^n\) and \(\tau^n\) are defined, then define \(\text{Net}^{n+1} := \text{Net}^n[\tau^n(4), \text{Net}^n(\tau^n(4)) + 1]\); \(\tau^{n+1} := \arg\min\{<\vec{v}>: \vec{v} \in \bigcup_{j=0}^{\infty} \vec{K}_j(\text{Net}^{n+1}(j))\}\).

The benefit of this search is that we do not have to calculate all the probabilities of all the possible values of Phi statistic - in doing so, it would take too much time. Net-type minimal search would keep on comparing the accumulated probabilities with the significance level. Only if those item values whose summation of probabilities are still under the significance level, then one needs to continue the search. Since all the values all are ordered in ascending order, one could effectively execute this search.

Definition 3.5. Define \(\Phi^1_i := \{\text{Ph}(\vec{v}): \vec{v} \in \tau^1\}\), where the Phi coefficient of \(\vec{v}\) is defined by \(\text{Ph}(\vec{v}) := \sqrt{\frac{N}{r_1 r_2}} \cdot \frac{c_2 \vec{v}(5) + c_1 \vec{v}(1) - c_1 c_2}{\sqrt{c_1 c_2}}\).

Based on Claim 6, we have the following definition.

Definition 3.6. Define \(\text{Prb}^i(p, N, r_1) = \{\frac{r_1}{\vec{v}(5)} \cdot \binom{N - r_1}{\vec{v}(1)} \cdot p^N \cdot (\frac{1}{p} - 1)^{\vec{v}(4)}: \vec{v} \in \tau^i\}\).

This set collects all the probabilities of the values in the minimal set \(\tau^i\). We could then accumulate all the probabilities in each \(\tau^i\) by the following set.

\(\text{Definition 3.7. Define } \text{AP}^n(p, N, r_1) := \sum_{k=1}^{n} \text{Prb}^k(p, N, r_1)\).

The accumulated probabilities in \(\text{AP}^n\) could then be used to test whether it has reached the threshold of a given significance level \(\alpha\).

Definition 3.8. (critical region) \(c = \arg\min\{n: \text{AP}^{n+1}(p, N, r_1) > \alpha \geq \text{AP}^n(p, N, r_1)\}\).

By this, we could calculate its critical region and do the statistical hypothesis testing. Now we summarize the whole algorithms as follows:

(i) Calculate each \(K_{S_{c_2}}\) for all \(0 \leq c_2 \leq \left[\frac{N}{2}\right]\).

(ii) Order each \(K_{S_{c_2}}\) and storer the result in the vector \(\vec{K}_{c_2}\) for all \(0 \leq c_2 \leq \left[\frac{N}{2}\right]\).

(iii) Test whether \(\text{AP}^n(p, N, r_1)\) has already reached the significance level \(\alpha\). If yes, then one could stop to do the statistical hypothesis testing; otherwise, one continues applying net-type minimal search to get the minimal sets \(\tau^n\).

(iv) Based on \(\tau^n\), one calculates \(\Phi^1_n\) and \(\text{Prb}^n(p, N, r_1)\).

(v) The process continues until \(\text{AP}^n\) reaches the significance level.
4. Properties

In this section, we derive all the properties that would justify the algorithms in the previous section.

Definition 4.1. (Inequality Order) If $\vec{v}, \vec{w} \in KS_{c_2}$, define

(i) $\vec{v} \prec \vec{w}$ iff $\vec{v}(2) < \vec{w}(2)$ or $[\vec{v}(2) = \vec{w}(2), \vec{v}(3) < \vec{w}(3)]$.

(ii) $\vec{v} \equiv \vec{w}$ iff $\vec{v}(2) = \vec{w}(2)$ and $\vec{v}(3) = \vec{w}(3)$.

(iii) $\vec{v} \preceq \vec{w}$ iff $\vec{v} \prec \vec{w}$ or $\vec{v} \equiv \vec{w}$.

Moreover, we use the notation $KS_{c_2}(i) \preceq KS_{c_2}(j)$ to denote that for all $\vec{v} \in KS_{c_2}(i), \vec{w} \in KS_{c_2}(j)$.

Lemma 4.1. For all $\vec{v}, \vec{w} \in KS_{c_2}(n)[\vec{v} \preceq \vec{w} \iff \vec{v}(3) \leq \vec{w}(3)]$.

Claim 8. $|\{k \in \mathbb{N} : m \cdot k < h\}| = \lfloor \frac{h-1}{m} \rfloor + 1$.

Indeed the size of each $KS_{c_2}(j)$ is determined by the following corollary.

Corollary 1. $|KS_{c_2}(j)| = \begin{cases} \lfloor \frac{c_1\cdot(j+1)-1}{c_2} \rfloor + 1, & \text{if } 0 \leq j \leq \min\{c_2, r_2\} \\ \lfloor \frac{c_1\cdot\min\{r_2, c_2\}-1}{c_2} \rfloor - \lfloor \frac{c_1\cdot(\min\{r_2, c_2\)}{c_2} \rfloor - 1, & \text{if } j > \min\{c_2, r_2\} \end{cases}$

Proof. It follows immediately from Claim 8 and the definition of $KS_{c_2}(j)$.

Lemma 4.2. $KS_{c_2}(n) \preceq KS_{c_2}(n+1)$ for any given $c_2$.

Theorem 4.3. (validity of algorithm) $c_2 \cdot [\min\{c_1, r_1\}] + c_1 \cdot [\min\{c_2, r_2\}] = \{c_1 \cdot \vec{v}(2) + \vec{v}(3) : \vec{v} \in KS_{c_2}\}$.

5. Conclusion

In this article, we have devised an approach to linearize the values of Kendall’s Phi statistic. This approach mainly exploits two stages of algorithms: Linearization of the values via each given $c_2$ and net-type minimal search algorithm. The combination of these two algorithms effectively construct the desired rejection region for the statistical hypothesis testing. This article would make Kendall’s Phi statistic much approachable and friendly, since these two algorithms could be further formalized in both analytical formula or computer programming. It would enhance the application of Kendall’s statistic in data analysis or other related contingency tables.

6. References

[1] Larry Wasserman. 2006. All of Nonparametric Statistics. New York: Springer.
[2] Jean Dickinson Gibbons and Subhabrata Chakraborti. 2003. Nonparametric Statistical Inferences. New York: Marcel Dekker.
[3] Charles J. Geyer and Glen D. Meeden. 2005. Rejoinder: Fuzzy and Randomized Confidence Intervals and P-Values. Statistical Science 20 (4): 384-387. doi:10.1214/088342305000000359.
[4] Christensen, Ronald. 1997. Log-linear models and logistic regression. Springer Texts in Statistics (Second ed.). New York: Springer-Verlag
[5] Everitt B.S. 2002. The Cambridge Dictionary of Statistics, CUP.
[6] Siegel Sidney. 1956. Non-parametric statistics for the behavioral sciences. New York: McGraw-Hill.

Acknowledgments

This work is supported by the Humanities and Social Science Research Planning Fund Project under the Ministry of Education of China (Grant No. 20XJAGAT001).