A New Weighted Rank Correlation

1Tahani A. Maturi and 2Ezz H. Abdelfattah
1Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia
2Faculty of Science, Helwan University, Cairo, Egypt

Abstract: Problem Statement: There have been many cases in real life where two independent sources have ranked n objects, with the interest focused on agreement in the top rankings. Spearman's rho and Kendall's tau coefficients assigned equal weights to all rankings. As a result, the literature proposed several weighted correlation coefficients with emphasis on the top rankings, including the top-down, weighted Kendall's Tau and Blest's correlation coefficient. Approach: This article introduced a new weighted rank correlation coefficient that was sensitive to agreement in the top rankings. It presented the limiting distribution under the null hypothesis of independence and provided a summary of quantiles of the exact null distribution for n = 3(1)9. Results: The article summarized the power comparison between the new weighted coefficient and other weighted coefficients, and showed that the new weighted rank correlation coefficient provided the locally most powerful rank test. Conclusions/Recommendations: The new weighted correlation should be used along with other weighted coefficients when the interest focused on agreement in the top rankings, in order to make an effective inference.

Key words: Weighted correlation, top ranking, top-down correlation coefficient, weighted Kendall's Tau, Blest's correlation coefficient.

INTRODUCTION

Every year many students want to apply for postgraduate courses and research, leading to a large number of applicants to universities. Postgraduate committee can choose only few of them, according to some criteria such as GPA and the average of their grades in the major courses that they have studied before. Since the number of the applicants is large, the aim is to minimize the effort and the cost of interviewing all the candidates while choosing the best among them. In such cases, a measure which gives more weight to those who have higher grades is required. Many other cases in life require more weight for values in the top in order to reach decision. For instance, a couple of panels of judges in one of the Olympic game wants to choose the best participants.

For such cases, correlation measures that give more weighted for the top rankings were presented by [1, 4, 5]. To review these measures briefly, let \( \{(X_i, Y_i), 1 \leq i \leq n\} \) be an independently and identically distributed (i.i.d.) sample from a bivariate distribution where \( q_i \) is the rank of \( Y \) whose corresponding \( X \) has rank \( i \) among \( \{X_i\} \). Throughout we assume that no ties occur among the variables being considered. If ties occur, the average of weighted score can be used. Iman and Conover [4] introduced the top-down correlation coefficient, \( R_t \), as:

\[
R_t = \frac{\sum_{i=1}^{n} S_i S_{q_i} - n}{n - S_t}
\]

(1)

Where, \( S_i \) is Savage score [4] defined as:

\[
S_i = \sum_{j=i}^{n} \frac{1}{j}
\]

Shieh [5] proposed the weighted Kendall's Tau, \( R_k \), which is given by:

\[
R_k = \frac{2}{m(m-1)} \sum_{i<j} \text{sgn}(i-j) \text{sgn}(q_i - q_j)
\]

(2)

Where, \( m \) is the number of top rankings taken into account and \( \text{sgn}(a) = -1, 0 \) or 1, if \( a < , = \) or > 0.

A graphical approach was proposed by [1], leading to a correlation coefficient \( R_b \), which is given by:

\[
R_b = \frac{2n+1}{n-1} - \frac{12}{n(n+1)^2(n-1)} \sum_{i=1}^{n} (n+1-i) q_i
\]

(3)
A new weighted rank correlation, $R_w$, that depends on weighted scores, will be introduced along with its asymptotic distribution under the null hypothesis of independence. Then some exact and approximated quantiles of $R_w$ are summarized. Power comparisons between $R_w$ and other reviewed coefficients will be presented. Finally, an example is given for illustration.

**MATERIALS AND METHODS**

**A new weighted rank correlation:** Let $(X_i, Y_i), (1 \leq i \leq n)$ be an i.i.d. sample from a bivariate distribution and let $(i, q_i), i = 1, 2, \ldots, n$, be paired rankings of $n$ objects, where $q_i$ is the rank of the $Y$ values whose corresponding $X$ has rank $i$ among all $\{X_j\}$. We define weighted scores as:

$$W_i = w^i$$  \hspace{1cm} (4)

Where, $i$ is the rank of the order observations in a sample of size $n$ and $0 < w < 1$.

The new weighted rank coefficient $R_w$ is obtained by computing the ordinary Pearson correlation coefficient, $r$, on the weighted scores,

$$R_w = \frac{\sum w_i q_i - a_i}{\sqrt{\sum w_i - n \cdot w_i}}$$  \hspace{1cm} (5)

Where, $a_i = w^i(1-w^i)^2 / (1-w^2)$ and $a_j = w^j(1-w^{2j}) / (1-w^2)$.

In another form, we can write $R_w$ as:

$$R_w = \sum_{i=1}^{n} \frac{w^i - n^{-1} \sqrt{a_i}}{\sqrt{w^i - n^{-1} \sqrt{a_i}}}$$  \hspace{1cm} (6)

The statistic $R_w$ has a maximum value of 1. However, its minimum possible value is only -1 for $n = 2$, similar as the top-down correlation \cite{4} and for $n \to \infty$ it increases from -1 towards approximately a value in the range from -2E-6 to -3E-4, depending on the value of $w$.

**The asymptotic distribution of $R_w$:** Now, the asymptotic distribution of $R_w$ is derived under $H_0$, the null hypothesis of independence. The alternative hypothesis of a positive dependence in the rankings can be detected using any of several statistics. The weighted rank correlation $R_w$ is more sensitive to agreement in the top ranks than to agreement in the bottom. For a test of $H_0$ that is equally sensitive to agreement among ranks at all levels, Spearman’s rho or Kendall’s tau correlation coefficient can be used. If the marginal distributions are normal and the alternative hypothesis is bivariate normal with positive correlation, the Pearson correlation coefficient, $\tau$, provides the most powerful test of $H_0$ against the alternative. Under $H_0$, the asymptotic distribution of $R_w$ is given by the following theorem:

**Theorem 1:** Under the null-hypothesis of independence, $E(R_w) = 0, V(R_w) = 1/(n-1)$ and the asymptotic distribution of $(n-1)^{1/2} R_w$ is the standard normal distribution.

**Proof:** The mean and the variance of the $R_w$, under $H_0$, are computed as follows. Since $E(\sum w_i q_i) = nE(w^i)E(w^q) = (1/n)w^i(1-w^q)^2 / (1-w^2) = a_i / n$, then by substituting in (5) we directly obtain that $E(R_w) = 0$. For the variance,

$$V(R_w) = \text{var}(R_w) = \frac{nV(w) - \text{cov}(w^i,w^j)\text{cov}(w^q,w^q)}{n(n-1)}$$

where

$$V(w) = V(w^q) = \frac{1}{n} \left( a_j - \frac{a_i}{n} \right)$$

and

$$\text{cov}(w^i,w^q) = \text{cov}(w^i,w^q) = E(w^i w^q) - E(w^i)E(w^q)$$

Therefore

$$V(\sum w_i q_i) = \frac{1}{n} \left( a_j - \frac{a_i}{n} \right) + \frac{1}{n(n-1)} \left( a_j - \frac{a_i}{n} \right)^2$$

and

$$\frac{1}{n-1} \left( a_j - \frac{a_i}{n} \right)$$

This completes the proof.
Using $a_i R_{w,f} = \frac{1}{n-1}(a_i - \bar{a})$ and $a_i = \frac{1}{n-1} \left( w_i - \bar{w} \right)$, we can write $R_w = \sum a_i R_{w,f} a_i R_{w,g}$. That is, $R_w$ is written as a linear rank statistic. Under $H_0$, using Theorem V.1.8 in Hájek and Šidák, the distribution of the statistic $R_w$ for $n \to \infty$ is asymptotically normal with mean 0 and variance $\sigma^2_{R_w} = V(R_w) = (n-1)^{-1}$.

Exact and approximate quantiles of $R_w$: When the null hypothesis is true, all permutations of ranks $(I, q)$, $1 \leq i \leq n$, are equally likely where $w$ can take any value between 0 and 1, exclusive. Then, exact and approximate quantiles of $R_w$ can be computed for chosen values of $w$, say 0.3, 0.6 and 0.9. Exact quantiles for $n = 3(1)9$ are summarized in Table 1 and for large $n$, approximate quantiles are shown in Table 2.

Table 1: Exact quantiles for the weighted correlation, $R_w$

| w   | n  | 0.90 | 0.95 | 0.975 | 0.99 | 0.995 | 0.999 |
|-----|----|------|------|-------|------|-------|-------|
| 0.3 | 3  | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 0.6 | 3  | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 0.9 | 3  | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

Fig. 1: Power curves of randomized tests of independence for $n=8$
RESULTS AND DISCUSSION

Numerical Example: To illustrate our new weighted rank correlation, we use a data set, in Table 4, that was also used by\(^5\). The data set considers two techniques, A and B, used to select the most effective variables out of 20 variables for evaluation of some software packages.

We see that the two techniques agree strongly on the top six variables. However, there is large disagreement between these techniques after that. In such circumstances, we may want to place more emphasis on the top rankings rather than equity over all ranking values. Therefore, we calculate some different weighted rank statistics, along with our weighted rank correlation at different weighted values. For each statistic the corresponding p-values are evaluated, these values are given in Table 5.

From Table 5 we can conclude that at different weight values, our weighted rank correlation and the top-down correlation provide strong evidence (p-value <0.001) against the null hypothesis of
independence of A and B. The criteria behind choosing
the weight depends on the degree of emphasis the user
may wish to apply to top ranks. However, we suggest
the weight \( w = 0.9 \) since as shown in Table 3, our new
weighted rank with \( w = 0.9 \) has higher power than other
rank correlation coefficients.

CONCLUSION

This article proposed a new weighted rank
correlation coefficient that was sensitive to agreement
in the top rankings. Under the null hypothesis of
independence, the proposed coefficient's limiting
distribution was derived along with the exact and
approximated quantiles for different sample sizes. As
shown, the test that depended on the new weighted rank
correlation coefficient was the locally most powerful
rank test. Therefore, when interest focused on the top
rankings, we recommended using the new weighted
rank correlation coefficient, together with other
weighted coefficients, to reach an effective decision.

A generalization of this article, when more than
two independent sources rank n objects with focus on
top rankings, known as a Concordance measure, will be
presented somewhere else.

REFERENCES

1. Blest, D.C., 2000. Rank correlation-An alternative
measure. Aust. N. Z. J. Stat., 42: 101-111. DOI:
10.1111/1467-842X.00110
2. Genest, C. and J. Plante, 2003. On Blest's measure
of rank correlation. Can. J. Stat., 31: 35-52.
http://direct.bl.uk/bld/PlaceOrder.do?UIN=132308
135&ETOC=RN&from=searchengine.
3. Hájek, J. and Z. Šidák, 1967. Theory of Rank
Tests. Academic Press, Inc. New York; London, pp
297, ISBN-10: 0123172500.
4. Iman, R.L. and W.J. Conover, 1987. A measure of
top-down correlation. Technometrics, 29: 351-357.
http://portal.acm.org/citation.cfm?id=37097.37105.
5. Shieh, G.S., 1998. A weighted Kendall's tau
statistic. Stat. Probability Lett., 39: 17-24. http://
cat.inist.fr/?aModele=afficheN&cpsidt=1618157.
6. Shieh G.S., Bai Z. and Tsai W. 2000. Rank Tests
for independence-with a weighted contamination
alternative. Stat. Sinica, 10: 577-593.
http://www3.stat.sinica.edu.tw/statistica/j10n2/j10n
212/j10n212.htm.