5-8-2020

Influence Function-based Empirical Likelihood Method for Kendall Rank Correlation Coefficient

Zhonglu Huang

INFLUENCE FUNCTION-BASED EMPIRICAL LIKELIHOOD METHOD FOR KENDALL RANK CORRELATION COEFFICIENT

by

ZHONGLU HUANG

Under the Direction of Gengsheng Qin, PhD

ABSTRACT

Correlation coefficients are used in statistics to measure the dependence between two variables. Kendall rank correlation coefficient is routinely used as a measure of association between two random variables in a number of circumstances in which the use of the Pearson correlation coefficient is inappropriate. In this thesis, we develop an influence function-based empirical likelihood interval for the Kendall rank correlation coefficient. Simulation studies are conducted to show good finite sample properties and robustness of the proposed method compared with existing methods. The proposed method is illustrated on a real UCLA graduate dataset.

INDEX WORDS: Confidence interval, Empirical likelihood, Kendall correlation coefficient

INFLUENCE FUNCTION-BASED EMPIRICAL LIKELIHOOD METHOD FOR KENDALL RANK CORRELATION COEFFICIENT

by

ZHONGLU HUANG

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2020

INFLUENCE FUNCTION-BASED EMPIRICAL LIKELIHOOD METHOD FOR KENDALL RANK CORRELATION COEFFICIENT

by

ZHONGLU HUANG

Committee Chair: Gengsheng Qin

Committee: Jun Kong

Jing Zhang

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

May 2020

ACKNOWLEDGEMENTS

First and foremost, I have to thank my parents for their love and support throughout my life. They kept me going on and this work would not have been possible without their input.

I wish to express my sincere appreciation to my advisor, Professor Gengsheng Qin, who has the substance of a genius: he convincingly guided and encouraged me to be professional and do the right thing even when the road got tough. Without his persistent help, the goal of this project would not have been realized.

I wish to thank the members of my dissertation committee: Dr. Jun Kong and Dr. Jing Zhang for generously offering their time, support, guidance and goodwill throughout the preparation and review of this document.

Finally, my deep and sincere gratitude to all of my friends and my manager and co-workers in AYS for their continuous and unparalleled help and support.

TABLE OF CONTENTS

LIST OF TABLES

1 INTRODUCTION

Correlation coefficient is a numerical measure of dependence between two variables. There are several types of correlation coefficients, each with their own definition and own range of usability and characteristics. For example, the Pearson correlation coefficient (Pearson, 1920) is a measure of the strength and direction of the linear relationship between two variables, the Spearman's rank correlation coefficient (Spearman, 1904) is a measure of how well the relationship between two variables can be described by a monotonic function, the Kendall rank correlation coefficient (Kendall, 1938) is used to measure the ordinal association between two measured quantities. Schaeffer and Levitt (1956) described a number of circumstances in which the use of the Pearson product-moment correlation is inappropriate and a rank order procedure is required, and discussed the advantages of Kendall rank correlation coefficient over Spearman's rank correlation coefficient. Croux and Dehon (2010) study robustness of the Kendall and Spearman correlations by means of their influence functions, and they found that the Kendall rank correlation is more robust and slightly more efficient than Spearman rank correlation.

Correlation measures are frequently used in applications. For instance, a correlation coefficient could be calculated to determine the level of correlation between the price of crude oil and the stock price of an oil-producing company. Since oil companies earn greater profits as oil prices rise, the correlation between the two variables is highly positive. In investing, a correlation is helpful in determining how well a mutual fund performs relative to its benchmark index, another fund or asset class. Also, a correlation statistic allows investors to determine when the correlation between two variables changes. Since loan rates are often calculated based on market interest rates, bank stocks always have a significant highly-positive correlation to interest rates.

In this thesis, we focus on the Kendall rank correlation coefficient (KCC). Let (X, Y) be a bivariate random vector with cumulative distribution function $H(x, y)$, the population KCC is defined as

$$R_K(H) \equiv E_H\left[sign\left((X_1 - X_2)(Y_1 - Y_2)\right)\right] = 2P_H[(X_1 - X_2)(Y_1 - Y_2) > 0], \quad (1.1)$$

where (X_1, Y_1) and (X_2, Y_2) are two independent copies of (X, Y).

For the bivariate normal distribution with population correlation coefficient ρ, denoted by Φ_ρ, we have (Blomqvist 1950),

$$R_K(\Phi_\rho) = \frac{2}{\pi} arcsin(\rho). \quad (1.2)$$

Let

$$\tilde{R}_K(H) = sin\left(\frac{1}{2}\pi R_K(H)\right). \quad (1.3)$$

Then $\tilde{R}_K(\Phi_\rho) = \rho$. $\tilde{R}_K(H)$ is called the Fisher consistent Kendall rank correlation coefficient (Maronna et al., 2006).

Let $(X_i, Y_i), i = 1, \dots, n$, be i.i.d. observations for (X, Y). Then the sample Kendall rank correlation coefficient is

$$r_K = \frac{2}{n(n-1)} \sum_{i<j} sign\left(\left(x_i - x_j\right)\left(y_i - y_j\right)\right). \quad (1.4)$$

r_K has range on [-1, 1], and is a consistent estimate for the population KCC $R_K(H)$. It measures the ordinal association between X_i's and Y_i's. If the agreement between the rankings of X_i's and Y_i's (i.e., the two rankings are the same) is perfect, $r_K = 1$. If the disagreement between the two rankings is perfect (i.e., one ranking is the reverse of the other), $r_K = -1$. If X and Y are independent, then r_K is approximately zero. r_K can be used as a test statistic for testing the statistical dependence of two variables if the sampling/asymptotic distribution of r_K can be found.

Croux and Dehon (2010) studied finite sample performances of several correlation estimators and showed that the sample KCC is resistant to outliers when contamination exists in the samples. The sample KCC is more robust and more efficient than the sample Spearman correlation coefficient. These advantages make the sample KCC a preferable estimator for $R_K(H)$. In this thesis, our goal is to propose a new non-parametric confidence interval for the KCC $R_K(H)$. The thesis is organized as follows. In section 2, we review a normal approximation-based interval for the KCC. In section 3, we propose a new influence function-based empirical likelihood interval for the KCC. In section 4, extensive simulation studies are conducted to examine the finite sample performance of the proposed interval. In section 5, a real example is used to illustrate the proposed method. In section 6, we conclude this thesis with a brief discussion. The proof of the main theorem is deferred to Appendix.

2 THE NORMAL APPROXIMATION METHOD

Assume that the bivariate random variable (X, Y) follows a distribution H. The influence function (IF) of a statistical functional R at a distribution H is defined as

$$IF\big((x,y),R,H\big) = \lim_{\varepsilon\to\infty} \frac{R\left((1-\varepsilon)H + \varepsilon\Delta_{(x,y)}\right) - R(H)}{\varepsilon}, \tag{2.1}$$

where $\Delta_{(x,y)}$ is a Dirac measure putting all its mass at (x, y) (see Hampel et al., 2011).

The influence function of the KCC is given as follows (Croux and Dehon, 2010):

$$IF\big((x,y),R_K,H\big) = 2\{2P_H[(X-x)(Y-y)>0] - 1 - R_K(H)\}. \tag{2.2}$$

Therefore, we have

$$\sqrt{n}\left(\frac{2}{n}\sum_{i=1}^{n} I\big[(X_i-\mu_x)\big(Y_i-\mu_y\big)>0\big] - 1 - R_K(H)\right) \xrightarrow{d} N(0,\sigma^2) \;\; as\; n\to\infty, \tag{2.3}$$

where $\mu_x = E(X)$, $\mu_y = E(Y)$, and

$$\sigma^2 = 4\mathrm{P_H}\big[(X-\mu_x)\big(Y-\mu_y\big)>0\big]\{1-\mathrm{P_H}\big[(X-\mu_x)\big(Y-\mu_y\big)>0\big]\}.$$

Using (2.3), a 100(1 − α) % normal asymptotic-based confidence interval for $R_K(H)$ can be constructed as

$$\frac{2}{\mathrm{n}}\sum_{i=1}^{n} I[(X_i-\bar{X})(Y_i-\bar{Y})>0] - 1 \pm \mathrm{Z}_{1-\frac{\alpha}{2}}\cdot\frac{\hat{\sigma}}{\sqrt{n}}. \tag{2.4}$$

where $\mathrm{z}_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the standard normal distribution, and $\hat{\sigma}$ is the square root of the estimator $\hat{\sigma}^2$ for σ^2defined by

$$\hat{\sigma}^2 = \frac{4}{\mathrm{n}}\sum_{i=1}^{n} I\big[(X_i-\mu_x)\big(Y_i-\mu_y\big)>0\big]\left\{1-\frac{1}{n}\sum_{i=1}^{n} I\big[(X_i-\mu_x)\big(Y_i-\mu_y\big)>0\big]\right\}. \tag{2.5}$$

3 EMPIRICAL LIKELIHOOD FOR THE KENDALL RANK CORRELATION COEFFICIENT

Hu, Jung and Qin (2020) recently proposed an influence function-based EL interval for the Pearson correlation coefficient. Motivated by their work, we proposed an influence function-based EL interval for the KCC $R_K(H)$ in this section.

From the influence function (2.2), it follows that the KCC $R_K(H)$ satisfies

$$E\left(2P_H\left[(X-\mu_x)(Y-\mu_y)>0\right]-1-R_K(H)\right)=0. \tag{3.1}$$

Let $W_i=(X_i,Y_i), i=1,\dots,n$, and $p=(p_1,\dots,p_n)$ be nonnegative numbers such that $\sum_{i=1}^n p_i=1$. From (3.1), an influence function-based EL for $R_K(H)$ can be defined as

$$L_0\big(R_K(H)\big)=\sup_p\left\{\prod_{i=1}^n p_i: p_i\geq 0, \sum_{i=1}^n p_i=1\,, \sum_{i=1}^n p_i\big(V(W_i)-R_K(H)\big)=0\right\}, \tag{3.2}$$

where

$$V_i(W_i)=2I\left[(X_i-\mu_x)\big(Y_i-\mu_y\big)>0\right]-1, i=1,\dots,n. \tag{3.3}$$

In practice, the population means $\left(\mu_x,\mu_y\right)$ are unknown. The sample means $(\bar{X},\bar{Y})$ can be used as the estimator of the population means. So, an influence function-based plug-in EL for $R_K(H)$ can be defined as

$$\hat{L}\big(R_K(H)\big)=\sup_p\left\{\prod_{i=1}^n p_i: p_i\geq 0, \sum_{i=1}^n p_i=1\,, \sum_{i=1}^n p_i\left(\hat{V}(W_i)-R_K(H)\right)=0\right\}, \tag{3.4}$$

where the pseudo sample

$$\hat{V}_i(W_i)=2I[(X_i-\bar{X})(Y_i-\bar{Y})>0]-1, i=1,\dots,n. \tag{3.5}$$

Using the Lagrange multiplier method, the expression for p_i can be obtained:

$$p_i=\frac{1}{n}\left\{1+\lambda\left(\hat{V}_i(W_i)-R_K(H)\right)\right\}^{-1}, i=1,\dots,n, \tag{3.6}$$

where λ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{V}_i(W_i)-R_K(H)}{1+\lambda\left(\hat{V}_i(W_i)-R_K(H)\right)}=0. \tag{3.7}$$

The corresponding influence function-based empirical log-likelihood ratio for $R_K(H)$ is

$$l\left(R_K(H)\right)=2\sum_{i=1}^{n}\log\left\{1+\lambda\left(\hat{V}_i(W_i)-R_K(H)\right)\right\}. \tag{3.8}$$

Theorem 1: If $E(X)$ and $E(Y)$ exist, and $R_K(H)$ is the true value of the Kendall rank correlation coefficient, then the asymptotic distribution of $l\left(R_K(H)\right)$ is a chi-square distribution with one degree of freedom, i.e.,

$$l\left(R_K(H)\right)\xrightarrow{d}\chi_1^2. \tag{3.9}$$

Based on the theorem above, a 100(1 − α) % influence function-based empirical likelihood (IEL) confidence interval for $R_K(H)$ can be constructed as

$$I_\alpha=\left\{\hat{R}_K(H): l\left(\hat{R}_K(H)\right)\le\chi_{1,1-\alpha}^2\right\}, \tag{3.10}$$

where $\chi_{1,1-\alpha}^2$ denotes the 100(1-α) % quantile of the chi-square distribution with one degree of freedom.

4 SIMULATION STUDIES

To examine the finite performance of the confidence interval (IEL) constructed by the influence function-based empirical likelihood method, we conduct a series of simulation studies. For comparison, the normal approximation-based confidence interval (NAI), and the Bootstrap confidence interval (Boot) for the KCC are also included in the studies.

We generate 5000 samples of three different sample sizes (n= 30, 50, 100) from several different distributions. And the correlation coefficient ρ is set to be 0.1, 0.5, and 0.9, respectively. The following underlying bivariate distributions are considered in the simulation studies.

a. Bivariate Normal Distributions

Scenario 1:

$$\binom{X}{Y} \sim N_2\left(\binom{0}{0} \cdot \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix}\right)$$

Scenario 2:

$$\binom{X}{Y} \sim N_2\left(\binom{1}{2} \cdot \begin{pmatrix} 2 & \sigma_{xy} \\ \sigma_{xy} & 3 \end{pmatrix}\right)$$

b. Bivariate mixed Normal Distributions

Scenario 3:

$$\binom{X}{Y} \sim 0.9N_2\left(\binom{0}{0} \cdot \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix}\right) + 0.1N_2\left(\binom{0}{1} \cdot \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix}\right)$$

which is a mixed normal distribution with 90% of observations from the first normal distribution and 10% of observations from the second normal distribution.

Scenario 4:

$$\binom{X}{Y} \sim 0.9N_2\left(\binom{0}{0} \cdot \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix}\right) + 0.1\log N_2\left(\binom{0}{1} \cdot \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix}\right)$$

where $\log N_2(\cdot,\cdot)$ is a log-normal distribution.

Scenario 5:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim 0.8N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix}\right) + 0.2N_2\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix}\right)$$

Scenario 6:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim (1-\epsilon) * N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix}\right) + \epsilon * N_2\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix}\right)$$

where $\epsilon = 0.01, 0.05, 0.1, 0.2$, respectively.

c. Bivariate t Distribution

Scenario 7:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim t_4\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix}\right)$$

d. Bivariate exponential Distribution

Scenario 8:

$$(X\ Y) \sim Biexp(\lambda_1, \lambda_2, \lambda_{12})$$

where $\lambda_1, \lambda_2, \lambda_{12}$ are parameters of the bivariate exponential distribution used in Marshall *et al*. (1967).

Simulation results are presented in Tables 1-8.

4.1 Bivariate normal distributions

From Tables 1-2, for bivariate normal distributions, most of the intervals have good coverage probabilities when ρ = 0.1 and 0.5. Compared with the existing intervals, the IEL method has good coverage probabilities under all the simulation settings except when ρ = 0.9 with a small sample size (n = 30). The IEL intervals perform equally well as the bootstrap method when the sample size larger than 50. The normal approximation-based intervals have

under-coverage problems when ρ = 0.1 and 0.9 with a small sample size. When ρ = 0.9 and n = 30, the IEL intervals and the normal approximation-based intervals have under-coverage problems with similar coverage probabilities while the bootstrap intervals have over-coverage problems. When the sample size increases, those problems have been corrected effectively.

Table 1 95% Confidence interval coverage probability and average length for $R_K(H)$ Scenario 1

		Coverage probability			Average length		
n	Method	0.1	0.5	0.9	0.1	0.5	0.9
30	IEL	0.956	0.949	0.932	0.680	0.645	0.512
	NAI	0.933	0.945	0.931	0.702	0.661	0.480
	Boot	0.955	0.956	0.976	0.521	0.470	0.290
50	IEL	0.955	0.950	0.961	0.537	0.509	0.401
	NAI	0.951	0.949	0.931	0.547	0.517	0.381
	Boot	0.943	0.954	0.969	0.389	0.348	0.203
100	IEL	0.944	0.940	0.955	0.386	0.364	0.275
	NAI	0.946	0.953	0.942	0.389	0.367	0.272
	Boot	0.947	0.946	0.951	0.268	0.237	0.131

Table 2 95% Confidence interval coverage probability and average length for $R_K(H)$ Scenario 2

		Coverage probability			Average length		
n	Method	0.1	0.5	0.9	0.1	0.5	0.9
30	IEL	0.956	0.949	0.931	0.680	0.648	0.515
	NAI	0.937	0.948	0.932	0.702	0.663	0.482
	Boot	0.949	0.957	0.974	0.522	0.471	0.290
50	IEL	0.955	0.951	0.960	0.537	0.508	0.401
	NAI	0.960	0.951	0.936	0.548	0.517	0.382
	Boot	0.952	0.956	0.966	0.390	0.349	0.202
100	IEL	0.946	0.942	0.958	0.386	0.364	0.276
	NAI	0.942	0.958	0.949	0.389	0.367	0.273
	Boot	0.944	0.950	0.955	0.268	0.238	0.131

4.2 Bivariate mixed Normal Distributions

For scenario 3-5, the underlying distributions are considered as bivariate normal distributions with different proportions of outliers. Table 3-5 list the results of these scenarios. Also, to test the robustness of the IEL method, scenario 6 will be used to estimate the confidence

interval of the Kendall correlation with a sample size n = 100 with different ratios of the outliers (ε = 0.01, 0.05, 0.1, 0.2).

For Table 3, we observe that the IEL method keeps good coverage probability when ρ = 0.1 and 0.5. For ρ = 0.9, the IEL intervals have a slightly over-coverage problem when n = 30 and 100. For the normal approximation-based intervals, a serious under-coverage problem exists compared with the IEL intervals. The bootstrap intervals perform a similar efficiency with the IEL intervals when ρ = 0.1 and 0.5 and have a worse over-coverage problem than the IEL intervals.

Table 3 95% Confidence interval coverage probability and average length for $R_K(H)$ Scenario 3

		Coverage probability			Average length		
n	Method	0.1	0.5	0.9	0.1	0.5	0.9
30	IEL	0.956	0.948	0.962	0.681	0.649	0.535
	NAI	0.934	0.946	0.890	0.702	0.665	0.513
	Boot	0.951	0.954	0.973	0.521	0.476	0.326
50	IEL	0.955	0.951	0.949	0.537	0.511	0.416
	NAI	0.954	0.929	0.921	0.548	0.519	0.403
	Boot	0.947	0.946	0.966	0.390	0.352	0.231
100	IEL	0.948	0.943	0.965	0.386	0.366	0.289
	NAI	0.942	0.938	0.966	0.389	0.369	0.288
	Boot	0.944	0.941	0.961	0.268	0.240	0.152

For Table 4, when outliers follow a bivariate log-normal distribution, the result will be different from scenario 3. The IEL intervals and the bootstrap intervals have good coverage probabilities when ρ = 0.1 and 0.5, while the normal approximation-based intervals have a serious under-coverage problem when the sample size is small. For ρ = 0.9 with a small sample size, the normal approximation-based interval is more efficient, and when the sample size becomes to 50, the IEL interval has the best performance in these three intervals.

Table 4 95% Confidence interval coverage probability and average length for $R_K(H)$ Scenario 4

		Coverage probability			Average length		
n	Method	0.1	0.5	0.9	0.1	0.5	0.9
30	IEL	0.948	0.952	0.919	0.672	0.632	0.523
	NAI	0.938	0.924	0.940	0.693	0.645	0.497
	Boot	0.965	0.960	0.982	0.518	0.451	0.280
50	IEL	0.951	0.945	0.943	0.530	0.496	0.409
	NAI	0.948	0.956	0.932	0.540	0.504	0.389
	Boot	0.965	0.956	0.966	0.388	0.332	0.194
100	IEL	0.957	0.959	0.929	0.381	0.355	0.281
	NAI	0.955	0.946	0.932	0.384	0.358	0.277
	Boot	0.966	0.952	0.957	0.266	0.225	0.124

When the ratio of outliers increases to 20%, the result shows in Table 5. the IEL interval keeps good coverage probability when $\rho = 0.1$ and 0.5. For $\rho = 0.9$, the IEL interval has a slightly over-coverage problem when n = 30. The normal approximation-based interval has a serious under-coverage problem when the sample size is small (n = 30 and 50). The bootstrap interval performs a similar efficiency, but the IEL interval is still the best method in these three methods.

Table 5 95% Confidence interval coverage probability and average length for $R_K(H)$ Scenario 5

		Coverage probability			Average length		
n	Method	0.1	0.5	0.9	0.1	0.5	0.9
30	IEL	0.959	0.942	0.972	0.681	0.650	0.554
	NAI	0.929	0.912	0.929	0.702	0.669	0.540
	Boot	0.947	0.952	0.977	0.522	0.480	0.343
50	IEL	0.955	0.960	0.941	0.538	0.513	0.431
	NAI	0.951	0.946	0.906	0.548	0.522	0.422
	Boot	0.955	0.948	0.969	0.390	0.354	0.245
100	IEL	0.943	0.958	0.954	0.386	0.368	0.301
	NAI	0.946	0.933	0.968	0.389	0.371	0.302
	Boot	0.946	0.949	0.966	0.268	0.243	0.162

To test the robustness of the IEL method, different proportions of outliers are applied in scenario 6, and the result shows in Table 6. All the intervals perform well in most of the conditions. Compared with the normal approximation-based interval and the bootstrap interval,

the IEL interval keeps a better efficiency. It won`t be influenced by the proportions of the outliers or the correlation coefficient.

Table 6 95% Confidence interval coverage probability and average length for $R_K(H)$ Scenario 6

n=100		Coverage probability			Average length		
ε	Method	0.1	0.5	0.9	0.1	0.5	0.9
0.01	IEL	0.948	0.941	0.952	0.386	0.365	0.278
	NAI	0.941	0.956	0.956	0.389	0.368	0.274
	Boot	0.946	0.945	0.960	0.268	0.238	0.134
0.05	IEL	0.943	0.945	0.947	0.386	0.366	0.282
	NAI	0.945	0.954	0.953	0.389	0.368	0.280
	Boot	0.950	0.944	0.925	0.267	0.239	0.144
0.1	IEL	0.944	0.938	0.958	0.386	0.366	0.289
	NAI	0.948	0.958	0.942	0.389	0.370	0.288
	Boot	0.944	0.945	0.960	0.268	0.240	0.152
0.2	IEL	0.949	0.950	0.945	0.386	0.367	0.301
	NAI	0.943	0.937	0.961	0.389	0.371	0.301
	Boot	0.949	0.948	0.959	0.268	0.242	0.161

4.3 Bivariate t Distribution

Table 7 95% Confidence interval coverage probability and average length for $R_K(H)$ Scenario 7

		Coverage probability			Average length		
n	Method	0.1	0.5	0.9	0.1	0.5	0.9
30	IEL	0.954	0.943	0.929	0.680	0.645	0.513
	NAI	0.929	0.940	0.933	0.702	0.661	0.482
	Boot	0.945	0.942	0.966	0.560	0.506	0.312
50	IEL	0.949	0.945	0.958	0.537	0.508	0.402
	NAI	0.951	0.949	0.933	0.547	0.517	0.382
	Boot	0.945	0.946	0.953	0.422	0.381	0.224
100	IEL	0.946	0.942	0.954	0.386	0.364	0.275
	NAI	0.945	0.956	0.953	0.389	0.368	0.272
	Boot	0.949	0.943	0.950	0.293	0.262	0.149

From Table 7, for bivariate t distributions, the IEL and bootstrap intervals perform well under all the simulation settings except n = 30 and ρ = 0.9, with an under-coverage problem for the IEL intervals and an over-coverage problem for bootstrap intervals. The IEL interval is more efficient than the normal approximation-based interval.

4.4 Bivariate exponential Distribution

Table 8 95% Confidence interval coverage probability and average length for $R_K(H)$ Scenario 8

		Coverage probability			Average length		
n	Method	0.1	0.5	0.9	0.1	0.5	0.9
30	IEL	0.931	0.917	0.936	0.674	0.591	0.391
	NAI	0.915	0.917	0.727	0.694	0.587	0.269
	Boot	0.946	0.949	0.893	0.541	0.539	0.529
50	IEL	0.940	0.938	0.937	0.531	0.457	0.286
	NAI	0.893	0.895	0.889	0.541	0.458	0.226
	Boot	0.948	0.951	0.954	0.409	0.409	0.353
100	IEL	0.903	0.914	0.906	0.381	0.322	0.207
	NAI	0.904	0.862	0.819	0.385	0.323	0.161
	Boot	0.944	0.949	0.959	0.282	0.285	0.191

The random sample of the bivariate exponential Distribution is generated by the algorithm proposed by Marshall and Olkin (1967):

For random samples $U_1 \sim exp(\lambda_1)$, $U_2 \sim exp(\lambda_2)$, $U_3 \sim exp(\lambda_{12})$, it follows

$\rho = \frac{\lambda_{12}}{\lambda_1 + \lambda_2 + \lambda_{12}}$, and the random sample (X, Y) with correlation coefficient ρ is

$$X = min(U_1, U_3), Y = min(U_2, U_3) \tag{4.1}$$

The simulation results for the bivariate exponential Distribution show in Table 8. The bootstrap intervals have good coverage probability under all the simulation settings except ρ = 0.9 and n = 30. While the IEL interval is the only one that keeps efficiency when ρ = 0.9 with a small sample size (n = 30). But under most of the simulation settings, the IEL interval has an under-coverage problem.

5 REAL DATA ANALYSIS

For the purpose of illustration, we apply the influence function-based empirical likelihood method to a UCLA graduate dataset (Acharya 2019). This dataset contains 400 cases and several parameters that are considered necessary during the application for master's Programs. We want to estimate the correlation coefficient between the GRE Scores and Chance of Admit. The sample Kendall rank correlation coefficient is 0.640. And the p-value of the Shapiro-Wilk multivariate normality test is 3.296e-07 (< 0.05), so we can reject the null (normality) hypothesis and consider the dataset as a non-normal data. Table 9 shows the results of the three methods. The 95% IEL confidence interval is (0.577, 0.725). The IEL method shows a similar efficiency compared with the asymptotical normality method and the bootstrap method. These results indicate a strong relationship between the GRE scores and the probability of admission.

Table 9 95% Confidence interval and length for the UCLA graduate admission data

	Confidence Interval	Length
IEL	(0.577, 0.725)	0.148
NAI	(0.581, 0.729)	0.148
Boot	(0.597, 0.677)	0.080

6 CONCLUSIONS

In this thesis, we develop a new influence function-based empirical likelihood method to construct a confidence interval for the Kendall rank correlation coefficient. The simulation study shows that this method performs well with underlying distribution being both the bivariate normal and nonnormal distributions. Also, when the outliers exist in the sample, the influence method performs good robustness. Based on the simulation study, we recommend the use of the IEL method when the underlying distribution is unknown or a nonnormal distribution with outliers.

REFERENCES

Fisher, Ronald A. "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population." Biometrika 10.4 (1915): 507-521.

Hampel, Frank R., et al. Robust statistics: the approach based on influence functions. Vol. 196. John Wiley & Sons, 2011.

Kendall, Maurice G. "A new measure of rank correlation." Biometrika 30.1/2 (1938): 81-93.

Blomqvist, Nils. "On a measure of dependence between two random variables." The Annals of Mathematical Statistics (1950): 593-600.

Croux, Christophe, and Catherine Dehon. "Influence functions of the Spearman and Kendall correlation measures." Statistical methods & applications 19.4 (2010): 497-515.

Hu, Xinjie, Aekyung Jung, and Gengsheng Qin. "Interval Estimation for the Correlation Coefficient." The American Statistician (2018): 1-8.

Maronna, R. A., R. Douglas Martin, and V. Y. Yohai. "Robust statistics: theory and practice." (2006).

Marshall, Albert W., and Ingram Olkin. "A multivariate exponential distribution." Journal of the American Statistical Association 62.317 (1967): 30-44.

Owen, Art. "Empirical likelihood ratio confidence regions." The Annals of Statistics (1990): 90

Pearson, Karl. "Notes on the history of correlation." Biometrika 13.1 (1920): 25-45. -120.

Schaeffer, Maurice S., and Eugene E. Levitt. "Concerning Kendall's tau, a nonparametric correlation coefficient." Psychological Bulletin 53.4 (1956): 338.

Spearman, Charles. "The proof and measurement of association between two things." The American journal of psychology 100.3/4 (1987): 441-471.

APPENDICES

Appendix A

We need the following Lemmas for the proof of Theorem 1.

Lemma 1: Under the conditions in Theorem 1, we have

$$\mathrm{n}^{-\frac{1}{2}}\sum_{i=1}^{n}\left(\hat{V}(W_i)-R_K(H)\right)\xrightarrow{\mathcal{L}}N(0,\sigma^2)$$

where $\sigma^2 = 4P_H\left[(X-\mu_x)\left(Y-\mu_y\right)>0\right]\left\{1-P_H\left[(X-\mu_x)\left(Y-\mu_y\right)>0\right]\right\}$.

Proof: From (3.5), we have the following decomposition:

$$\mathrm{n}^{-\frac{1}{2}}\sum_{i=1}^{n}\left(\hat{V}(W_i)-R_K(H)\right)=\mathrm{n}^{-\frac{1}{2}}\sum_{i=1}^{n}\left(\hat{V}(W_i)-V(W_i)\right)+\mathrm{n}^{-\frac{1}{2}}\sum_{i=1}^{n}\left(V(W_i)-R_K(H)\right)$$

$$\equiv I_1+I_2. \qquad (A.1)$$

From $-\infty < E(X)=\mu_x<\infty, -\infty<E(Y)=\mu_y<\infty, \bar{X}=\mu_x+o(1)\ a.s. and\ \bar{Y}=\mu_y+o(1)\ a.s.$, it follows that

$$\hat{V}(W_i)-V(W_i)=2\left(I[(X_i-\bar{X})(Y_i-\bar{Y})>0]-I\left[(X_i-\mu_x)\left(Y_i-\mu_y\right)>0\right]\right)$$
$$=0\ a.s. for\ \forall\ i, as\ n\to\infty. \qquad (A.2)$$

Hence,

$$I_1=0\ a.s. \qquad (A.3)$$

From (2.3), we have

$$I_2\xrightarrow{\mathcal{L}}N(0,\sigma^2). \qquad (A.4)$$

Lemma 1 follows from (A.1), (A.3) and (A.4) right away.

Lemma 2: Under the conditions in Theorem 1, we have that

$$\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(\hat{V}(W_i)-R_K(H)\right)^2\xrightarrow{p}\sigma^2.$$

Proof: From $|V(W_i)-R_K(H)|\le 5$ and (A.2), it follows that

$$\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(\hat{V}(W_i)-R_K(H)\right)^2=\mathrm{n}^{-\frac{1}{2}}\sum_{i=1}^{n}\left(V(W_i)-R_K(H)\right)^2+\mathrm{n}^{-\frac{1}{2}}\sum_{i=1}^{n}\left(\hat{V}(W_i)-V(W_i)\right)^2$$

$$+\frac{2}{n}\sum_{i=1}^{n}\left(\hat{V}(W_i)-V(W_i)\right)\left(V(W_i)-R_K(H)\right)$$

$$=E\left(V(W_i)-R_K(H)\right)^2+o_p(1)$$

$$=\sigma^2+o_p(1)$$

The Proof of Theorem 1.

Using similar arguments in Owen (1990), we can prove that $\lambda=O_p(n^{-0.5})$. Applying Taylor`s expansion, we obtain that

$$l\left(R_K(H)\right)=2\sum_{i=1}^{n}log\left[1+\lambda\left(\hat{V}(W_i)-R_K(H)\right)\right]$$

$$=2\sum_{i=1}^{n}\left[\lambda\left(\hat{V}(W_i)-R_K(H)\right)-\frac{1}{2}\left(\lambda\left(\hat{V}(W_i)-R_K(H)\right)\right)^2\right]+r_n$$

where

$$|r_n|\le C\sum_{i=1}^{n}\left|\lambda\left(\hat{V}(W_i)-R_K(H)\right)\right|^3\le C|\lambda|^3n=O_p(n^{-0.5})$$

From (3.7), if follows that

$$\lambda=\frac{\sum_{i=1}^{n}\left(\hat{V}(W_i)-R_K(H)\right)}{\sum_{i=1}^{n}\left(\hat{V}(W_i)-R_K(H)\right)^2}+O_p(n^{-0.5})$$

$$\sum_{i=1}^{n}\lambda\left(\hat{V}(W_i)-R_K(H)\right)=\sum_{i=1}^{n}\left(\lambda\left(\hat{V}(W_i)-R_K(H)\right)\right)^2+O_p(1)$$

Therefore, by Lemmas 1-2, we have that

$$l\left(R_K(H)\right) = \sum_{i=1}^{n} \lambda \left(\hat{V}(W_i) - R_K(H)\right) + o_p(1) = \frac{\left[\sum_{i=1}^{n}\left(\hat{V}(W_i) - R_K(H)\right)\right]^2}{\sum_{i=1}^{n}\left(\hat{V}(W_i) - R_K(H)\right)^2} + o_p(1) \xrightarrow{\mathcal{L}} \chi_1^2$$

The proof of the Theorem 1 is thus completed.