American Academic Scientific Research Journal for Engineering, Technology, and Sciences
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402
http://asrjetsjournal.org/

Which Correlation Coefficient Should Be Used for Investigating Relations between Quantitative Variables?

Ebru Temizhan[a], Hamit Mirtagioglu[b], Mehmet Mendes[c*]

[a,c]*Canakkale Onsekiz Mart University, Agriculture Faculty, Biometry and Genetics Unit, 17100, Canakkale, Turkey*

[b]*Bitlis Eren University, Faculty of Arts and Sciences, Department of Statistics, Bitlis, Turkey*

[a]*Email: ebrutemizhan@gmail.com,* [b]*Email: hmirtagioglu@beu.edu.tr,* [c]*Email: mmendes@comu.edu.tr*

Abstract

Since the purpose of many studies is to describe and summarize the relations between two or more variables, the correlation analysis has been become one of the most fundamental statistical concepts for many researchers. There are different correlation coefficients have been developed and proposed for different cases. In this stage, it is extremely important to aware of which correlation coefficient(s) is more appropriate to use based on the measurement levels, type of the variables, distribution of the variables, type of relations between the variables, and presence of outliers or not in dataset. In this study, nine different correlation coefficients have been compared in terms of Type I error rate and test power under different experimental conditions. As a result, it has been possible to produce information about which correlation coefficient is more appropriate to use in which situations. Results of this simulation study showed that the performances of these correlation coefficients are affected by sample size and effect size rather than the distribution shape. When both the type I error and test power estimates are evaluated together, the Pearson's correlation, Winsorized, Spearman Rank, and Kendall-Tau correlation coefficients are seem to be the most appropriate coefficients for many experimental conditions.

Keyword: correlation coefficient; type I error; test power; simulation; robust methods.

1. Introduction

Since many studies carried out in practice were conducted to investigate the relationships between variables, the correlation has become one of the most basic terms. As a result, correlation analysis has become one of the most widely used statistical techniques in all branches of science [1, 2, 3, 4]. Although many different correlation coefficients have been proposed in literature (i.e. Pearson, Spearman, Kendall Tau, Winsorized, Permutation-based, Distance, Sheapard, Percentage-bend, Blomquvist, etc.) the Pearson Product-Moment correlation is the most commonly used by scientists and researchers, despite its lack of robustness [4, 5, 6].

---

* Corresponding author.

However, it should also be kept in mind that the Pearson correlation coefficient can only capture the linear relationship between normally distributed variables. Pearson correlation also requires some other assumptions namely no outliers in data set, and a homoscedasticity between the variables of interest. Therefore, the reliability of this correlation is dependent on whether or not these assumptions are met as well as an adequate sample size ($n \geq 10$) [4, 7, 8, 9]. However, in practice, situations in which at least one of these assumptions is not fulfilled are quite common that limits the use of the Pearson correlation. One of the other conditions that limits the use of Pearson correlation coefficient is differences in the way of collecting data based on the aim of the study. Since there are so many correlation coefficients have been proposed and many of them are still not included in the basic statistical text books and statistical package programs, answering of the question of which coefficient is the most appropriate for the dataset studied is extremely important. It is because that way it will be possible to reveal the relationships that actually exist between variables in correct way.

This study was mainly conducted for two purposes. The first aim of this study is to introduce the nine correlation coefficients that are most likely to be used in practice, to indicate the differences between them, and to explain which correlation coefficient is more appropriate to use in which situations. The second purpose is to compare the performances of the correlation coefficients which can be used for the same purpose based on Monte Carlo Simulation Study.

2. Material and Methods

The materials of this study have been consisted of simulated data from bivariate normal and bivariate lognormal distributions under different variance-covariance structures and sample sizes.

2.1. Simulation study

In this study, the Pearson, Spearman's Rank, Kendall's Tau, Percentage Bend, Winsorized, Distance, Biweight Midcorrelation, Blomqvist's, and Hoeffding's D correlation coefficients have been compared with respect to type I error rate (α) and test power ($1-\beta$). For this aim, bivariate normal and bivariate lognormal variables have been generated for different sample size combinations (n=10, 30, 100, and 500) and true population correlations or effect sizes (ρ=0.0, 0.30, 0.60, and 0.90). All computations were performed by using R software [10]; R Studio [11].

2.2. Correlation coefficients

Although many different correlation coefficients have been developed and proposed, it has only been focused on commonly used correlations or will promising correlations in the future.

2.2.1. Pearson correlation coefficient

Pearson correlation coefficient is the most widely used correlation and perhaps the most well known correlation by researchers and scientists to measure the linear relationship between two variables. The formula for the Pearson correlation coefficient is *(1)*

$$r_{xy} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2 \sum(Y_i-\bar{Y})^2}} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}} \qquad (1)$$

Where $\sum d_x d_y$ product sum of squares,

$\overline{X}$ and $\overline{Y}$ are the sample means of X and Y, and

$\sum d_x^2$ and $\sum d_y^2$ sum of squares of X and Y, respectively [3, 4, 9].

2.2.2. Spearman rank correlation

The Spearman rank correlation test is basically the nonparametric version of the Pearson correlation coefficient and it provides to investigate the linear relations between two variables. This correlation coefficient can also adapt to ordinal data. Since it is a nonparametric coefficient, it will be appropriate to use especially when the data have violated parametric assumptions (i.e. non-normally distributed data), sample size is small and there is an outlier problem in data set. It is possible to interpret this correlation in terms of explained variability of the ranks. It can also be used to assess the monotonic relations based on the rank of the observations. This is an important issue because linear relationships are monotonic, but all monotonic relationships are not needed to be linear. Although it is a nonparametric coefficient or distribution free test the Spearman rank correlation coefficient requires a few assumptions. The assumptions of this coefficient are that as the data must be at least ordinal and the scores on one variable must be monotonically related to the other variable. The Spearman rank correlation can be computed by using following formulas *(2, 3)*:

$$r_{sr} = \frac{\sum(R_{ix}-\bar{R}_x)(R_{iy}-\bar{R}_y)}{\sqrt{\sum(R_{ix}-\bar{R}_x)^2 \sum(R_{iy}-\bar{R}_y)^2}} \qquad (2) \text{ or}$$

$$r_{sr} = \frac{1-6\sum d_i}{n(n^2-1)} \qquad (3)$$

Where R_{ix} and R_{iy} are the ranks of the ith X and Y values.

$\overline{R}_x$ and $\overline{R}_y$ are the means of the R_{ix} and R_{iy} values,

$d_i = X_i - Y_i$ is the difference between the ranks of corresponding variables,

N is the number of observations [3, 4, 12, 13, 14, 15].

2.2.3. Kendall's tau correlation coefficient

As in the Spearman rank correlation, the Kendall's Tau correlation coefficient is a nonparametric measure of association and it is used to evaluate the relationship between two ordinal variables. This coefficient is based on the number of concordances and discordances in paired observations. When paired observations vary together the concordance occur, while discordance occurs when paired observations vary differently. Therefore,

conceptually, Kendall's tau correlation is used for assessing the proportion of discrepancy between concordant and discordant pairs. Kendall and Gibbons (1990) reported that any two pairs of rank (X_i, Y_i) and (X_j, Y_j) are concordant if $Y_i < Y_j$ when $X_i < X_j$ or if $Y_i > Y_j$ when $X_i > X_j$ or if $(X_i-Y_i)(X_j-Y_j)>0$ [16]. And, any two pairs of rank (X_i, Y_i) and (X_j, Y_j) are discordant if $Y_i < Y_j$ when $X_i > X_j$ or if $Y_i > Y_j$ when $X_i < X_j$ or if $(X_i-Y_i)(X_j-Y_j)<0$.

Let C be the number of concordant pairs, D be the number of discordant pairs, and n be the sample size, in this case, based on n subjects to be ranked, there will be k=n(n-1)/2 possible comparisons between any pairs of rank (X_i, Y_i) and (X_j, Y_j). Based on this information, the Kendall's Tau correlation can be computed by using following formula *(4)*.

$$r_{tau} = \frac{\neq concordant\ pairs - \neq discordant\ pairs}{\frac{n(n-1)}{2}} \qquad (4)$$

This correlation varies between -1 and +1 [4, 17, 18].

2.2.4. Winsorized correlation

Winsorized correlation is recommended especially for the cases where outliers presence in datasets. This correlation is another robust alternative of the Pearson correlation in case of outlier exists. The computation of Winsorized correlation is quite simple. It uses Person's correlation formula applied on the Winsorized data. Winsorized correlation coefficient, which is computed after the k smallest observations are replaced by the (k+1)st smallest observation, and the k largest observations are replaced by the (k+1)st largest observation. Therefore, the observations are winsorized at each end of both X and Y [4, 19, 20]

Let (X_1,Y_1), (X_2,Y_2),…,(X_n,Y_n) be a random sample from any bivariate distribution. We winsorize X_i to W_i and Y_i to V_i any bivariate distribution as described above. Then, the Winsorized correlation coefficient, r_w, is computed same as the Pearson moment correlation using the Winsorized data as follows *(5)*:

$$r_w = \frac{\sum(W_i-\bar{X}_w)(V_i-\bar{Y}_w)}{\sqrt{\sum(W_i-\bar{X}_w)^2\sum(V_i-\bar{Y}_w)^2}} \qquad (5)$$

where $\bar{X}_w$ and $\bar{Y}_w$ are the Winsorized means of X and Y variables, respectively. Let γ denote the Winsorized percent and define g =γn; then, r_w is distributed as t-distribution with (n-2g-2) d.f [4, 19].

2.2.5. Hoeffding's measure of dependence, D

Hoeffding's measure of dependence, D, is a nonparametric measure of association that detects more general departures from independence. The statistic approximates a weighted sum over observations of chi-square statistics for two-by-two classification tables [13]. Each set of (X,Y) values are cut points for the classification. The formula for Hoeffding's D is *(6)*

$$D = 30\frac{(n-2)(n-3)D_1 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)} \qquad (6)$$

Where $D_1 = \sum_i (Q_i - 1)(Q_i - 2)$

$$D_2 = \sum_i (R_{ix} - 1)(R_{ix} - 2)(R_{iy} - 1)(R_{iy} - 2)$$

$$D_3 = \sum_i (R_{ix} - 1)(R_{iy} - 2)(Q_i - 1)$$

Where R_{ix} and R_{iy} are the ranks of the ith X and Y values. Q_i is also known as bivariate rank and it represents the number of points with both X and Y values less than the i[th] point. A point that is tied on only the X value or Y value contributes ½ to Q_i if the other value is less than the corresponding value for the i[th] point. A point that is tied on both the X and Y valuee contribute 1/4 to Q_i.

The D statistic values are between -0.5 and 1 (1 indicating complete dependence) when there is no tie in dataset. However, when ties occur, the D statistic may result in a smaller value. That is, for a pair of variables with identical values, the Hoeffding's D statistic may be less than 1. With a large number of ties in a small data set, the D statistic may be less than -0.5. For more information on Hoeffding's D [22, 23, 24].

2.2.6. Distance correlation

It is well known that the relations between the variables are not always linear. In some cases, relations between the variables are nonlinear. Therefore, a correlation coefficient that also provides to detect nonlinear relationships between variables is needed. For this purpose, distance correlation coefficient has been introduced by Szekely and his colleagues 2007, 2009 and Szekely & Rizzo 2012 and 2013 [25, 26, 27, 28]. Since the distance correlation can applicable to random variables of any dimension and measures both linear and non-linear association between two random variables or random vectors, it provides more information when compared to the Pearson's correlation [34]. Due to such advantages of the distance correlation it has been become increasingly used in many field of sciences [29, 30, 31, 32]. In contrast to the Pearson's correlation, it equals to zero if and only if the variables are independent.

Therefore, the distance correlation provides more information than the Pearson's correlation coefficient, and the number of references to the distance correlation method has increased rapidly across a wide variety of fields [30, 31, 32].

2.2.7. Biweight midcorrelation

It is well known that the Pearson correlation is very sensitive to outliers. Since the Biweight midcorrelation is median-based coefficient it is less sensitive to outliers and thus it can be used as a robust counterpart to the Pearson's correlation [33, 34]. Biweight midcorrelation of two numeric vectors X=($x_1, x_2, \dots, x_m$) and Y=($y_1, y_2, \dots, y_m$) is defined as *(7)*

$$Bicorr(X,Y) = \frac{\sum_{i=1}^{m}(X_i - Med(X))w_i^{(X)}(Y_i - Med(Y))w_i^{(Y)}}{\sqrt{\sum_{j=1}^{n}[(Y_i - Med(Y))w_i^{(X)}]^2 \sum_{k=1}^{m}[(Y_k - Med(Y))w_k^{(Y)}]^2}} \qquad (7)$$

Where $w_i^{(X)}$ stand for weight for Xi and defined as *(8)*

$$w_i^{(X)} = (1 - u_i^2)^2 I(1 - |u_i|) \qquad (8)$$

Where the indicator $I(1 - |u_i|)$ take 1 if $1 - |u_i| > 0$ and 0 otherwise.

Where Med(X) is the median of X and Med(Y) is the median of Y.

u_i *(9)* and v_i *(10)* are respectively defined as follow:

$$u_i = \frac{X_i - Med(X)}{9AMed(X)} \qquad (9)$$

$$v_i = \frac{Y_i - Med(Y)}{9AMed(Y)} \qquad (10)$$

Where AMed (X) and AMed(Y) are the median absolute deviation of X and Y *(11, 12)* [34, 35, 36].

$$AMed(X) = Med(|X_i - Med(X)|) \qquad (11)$$

$$AMed(Y) = Med(|Y_i - Med(Y)|) \qquad (12)$$

2.2.8. Blomqvist's coefficient

Blomqvist's coefficient or Blomqvist's Beta / medial correlation is one of the other median-based nonparametric correlation coefficients. This coefficient has several advantages over Spearman's or Kendall's [37, 38]. Blomqvist (1950) suggested the following formula for this coefficient *(13)* [37]:

$$\widehat{\beta_n} = \frac{2n_1}{n_1 + n_2} - 1 \qquad (13)$$

For a pair of continuous variables X and Y, the Blomqvist's β can be computed *as* $\beta = \{(X - \bar{x})(Y - \bar{y}) > 0\} - \{(X - \bar{x})(Y - \bar{y}) < 0\}$

Where $\bar{x}$ and $\bar{y}$ are the median of X and Y, respectively [38].

2.2.9. Percentage bend correlation coefficient

The percentage bend correlation (ρ_{pb}) is a robust alternative to Pearson's correlation [8]. The percentage bend correlation estimator is both resistant and robust of efficiency. Although when the underlying data are bivariate normal, ρ_{pb} gives essentially the same values as Pearson's correlation, this correlation is more robust in slightly changes in data that Pearson's correlation in general. The ρ_{pb} belongs to class of correlation measures which protect against marginal distribution (*X* and *Y*) outliers. Therefore, this correlation is similar to Spearman's Rho, Kendall's Tau, and biweight midcovariance correlation. The percentage bend correlation between variables X and Y is computed as following computational steps given below:

Step 1: Set $m=(1-\beta)n)+0.5$ and round m down to the nearest integer

Step 2: Let $W_i = |X_i - M_x|$ for $i = 1, 2, 3, \dots, n$ where M_x is the median of X

Step 3: Sort the W_i in ascending order

Step 4: $\widehat{W}_x = W(m)$ (i.e., the m-th order statistic). $W(m)$ is the estimation of the $(1-\beta)$ quantile of W.

Step 5: Sort X values. Compute the number of values of $(X_i - M_x)/\widehat{W}_x(\beta)$ that are less than -1 and the number that are greater than +1 and store in i1 and i2 respectively. Then it is calculated the terms below respectively *(14, 15, 16)*.

$$S_x = \sum_{i=i1+1}^{n-i2} X_i \qquad (14)$$

$$\widehat{\emptyset}_x = \frac{\widehat{W}_x(i2-i1)+S_x}{n-i1-i2} \qquad (15)$$

$$U_i = \frac{X_i-\widehat{\emptyset}_x}{\widehat{W}_x} \qquad (16)$$

Step 6: Repeat the above calculations on the Y variable. Store corresponding quantities in $\widehat{W}_y$, $\widehat{\emptyset}_y$, and V_i.

Step 7: Define the function *(17)*

$$\gamma(X) = max[-1, min(1, x)] \qquad (17)$$

Step 8: Compute the following terms *(18, 19)*

$$A_i = \gamma_i(U_i) \qquad (18)$$

$$B_i = \gamma_i(V_i) \qquad (19)$$

Step 9: Compute the percentage bend correlation coefficient as below *(20)*:

$$r_{pbn} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}} \qquad (20)$$

The value of β is selected between 0 and 0.5. Higher values of β result in a higher breakdown point at the expense of lower efficiency [39, 40, 41].

3. Results and Discussion

Since the purpose of many studies is to describe and summarize the relations between two or more variables, the correlation analysis is widely used by the researchers in practice [4, 7, 8, 42, 43, 44, 45]. As a result, the

correlation analysis has been one of the most fundamental statistical concepts used in almost all branches of sciences to assess relations between two or more variables. Although correlation analysis is one of the most widely used statistical techniques in practice, it is worth noting that the researchers (especially non-statisticians) have a difficulty in determining an appropriate coefficient for their dataset due to a large number of correlation coefficients have been developed and proposed. Main reason is that the majority of the proposed correlation coefficients (although it is possible to find information about them in different papers) is not included yet in statistical text books and commonly used statistical package programs (i.e. SPSS, Minitab, Statistica, SAS, etc). Since the correlation coefficients which will be able to use to evaluate the relation between the variables varies based different experimental factors like the measurement levels of the variables, type of the variables, it is extremely important to aware of using an appropriate correlation in assessing the relations between the variables. It is because; the use of various correlation coefficients for the same set of data may lead to significantly different conclusions.

As a result, the researchers are confused about which results are reliable. In this case, the following questions might pop up on the researchers' mind: Why there are so many correlations? What are the differences among them? Which one(s) is more appropriate for investigating the relations between pair of variables? How to determine the most appropriate correlation(s) for our data set? Which criteria should be considered in determining the most appropriate on? How do we calculate? In this study, nine of correlation coefficients namely the Pearson, Spearman's Rank, Kendall's Tau, Percentage Bend, Winsorized, Distance, Biweight Midcorrelation, Blomqvist's, and Hoeffding's D correlation coefficients have been introduced then a comprehensive simulation study has been carried out for comparing performances of these tests. The Type I error rate and test power estimates have been used as performance criterias. Simulation study results have been presented in Table 1.

Table 1: Means, standart errors, type I error rates and test powers when samples taken from μ=5:5, σ=1:1 bivariate normal and lognormal distributions

		Bivariate Normal				**Bivariate Lognormal**			
n	**Correlation**	**ρ=0**	**ρ=0.30**	**ρ=0.60**	**ρ=0.90**	**ρ=0**	**ρ=0.30**	**ρ=0.60**	**ρ=0.90**
		α	**1-β**	**1-β**	**1-β**	**α**	**1-β**	**1-β**	**1-β**
10	Pearson	4.92	12.56	47.25	98.29	5.91	16.25	41.31	93.57
	Kendall	4.98	10.47	37.51	93.36	4.98	10.37	37.67	93.5
	Spearman	5.77	11.91	40.45	94.18	5.77	12.06	40.37	94.43
	Winsorize	4.92	12.56	47.25	98.29	5.91	16.25	41.31	93.57
	Distance	7.66	14.81	45.62	95.79	7.59	14.56	40.61	94.09
	Biweight	0.32	0.67	2.60	5.53	0.21	0.56	1.45	4.69
	Hoeffding	0.68	1.11	2.73	5.51	0.68	1.09	2.96	5.47
	Percentage	0.30	0.85	2.79	5.67	0.28	0.85	2.51	5.7
	Blomqvist	0.05	0.06	0.49	2.07	0.05	0.1	0.42	2.01
30	Pearson	5.31	37.64	95.54	100	5.5	27	76.12	100
	Kendall	5.13	32.72	92.5	100	5.13	33.22	92.2	100

	Spearman	5.35	33.49	92.83	100	5.35	33.69	92.48	100
	Winsorize	5.33	36.15	94.83	100	4.96	29.63	84.93	100
	Distance	7.13	36.68	93.67	100	6.73	30.92	88.91	100
	Biweight	4.69	34.17	93.79	100	4.59	21.87	72.58	99.99
	Hoeffding	6.76	32.73	89.59	100	6.76	33.2	89.91	100
	Percentage	4.87	33.54	92.69	100	4.97	31.39	89.78	100
	Blomqvist	2.85	12.33	52.21	98.57	2.85	12.46	52.68	98.48
100	Pearson	4.76	86.62	100	100	4.43	52.97	99.37	100
	Kendall	4.70	82.78	100	100	4.7	82.47	100	100
	Spearman	4.83	83.07	100	100	4.83	82.65	100	100
	Winsorize	4.73	84.88	100	100	4.59	70.62	100	100
	Distance	6.58	84.29	100	100	6.22	74.55	100	100
	Biweight	5.06	85.39	100	100	5.36	62.33	99.89	100
	Hoeffding	5.44	78.67	99.99	100	5.44	78.41	99.99	100
	Percentage	5.08	82.73	100	100	5.15	77.25	100	100
	Blomqvist	7.14	56.14	99.31	100	7.14	55.92	99.32	100
500	Pearson	4.99	100	100	100	4.37	98.26	100	100
	Kendall	5.08	100	100	100	5.08	100	100	100
	Spearman	5.12	100	100	100	5.12	100	100	100
	Winsorize	5.15	100	100	100	5.03	100	100	100
	Distance	6.76	100	100	100	6.4	99.99	100	100
	Biweight	4.76	100	100	100	10.31	99.92	100	100
	Hoeffding	4.96	100	100	100	4.96	100	100	100
	Percentage	4.94	100	100	100	5.07	100	100	100
	Blomqvist	5.82	99.21	100	100	5.82	99.29	100	100

Results of this simulation study showed that the performances of these correlation coefficients are affected by sample size and effect size rather than the distribution shape. When the type I error estimates of these correlation coefficients are evaluated, it is clearly seen that the Pearson's correlation, Winsorized, Spearman Rank, and Kendal-Tau correlation coefficients gave the most stable results in terms of protecting the Type I error rates at 5.00 % level regardless of sample size and distribution shape. The type 1 error estimations of these correlations changed between 4.71 and 5.77 % in general. All these estimations are fall between both Bradly's (1978) liberal criteria ($0.025 \leq \alpha^{*} \leq 0.075$) and Cochran's (1954) criteria ($0.045 \leq \alpha^{*} \leq 0.055$) [46, 47]. On the other hand, the Distance, Biweight mid, Hoeffding, Percentage Bend and Blomqvist correlation coefficients are the most affected coefficients from the sample size regardless of distribution shape. Type I error estimates of these coefficients varied between 0.05 and 7.66% when samples were taken from bivariate normal populations while they changed between 0.05 and 10.31% when samples were taken from bivariate lognormal populations. However, in parallel with the increase in sample size, it has been observed that the type 1 error estimates of

these correlation coefficients tend to gradually approach to 5.00 %. When n=500, all tests (relatively except for the Distance and Blomqvist correlations) gave similar results in terms of keeping the type I error rate at 5.00 % level. Under these experimental conditions, all correlation coefficients kept the type 1 error rate at 5.00% level. Although some small differences have been observed in type I error estimates of these correlations for the bivariate normal and bivariate lognormal distributions, the type I error estimates for both distributions are generally similar. Thus, it is possible to conclude that the distribution shape is not the key factor that affect the type I error estimates.

When test power estimates of these correlations are evaluated, all coefficients produced very low test power values (0.05 and 37.64%) when sample size is between 10 and 30 for rho=0.30. However, as the sample size and effect size increase the test power values increase as well for all coefficients regardless of distribution shape. On the other hand, when all experimental conditions are evaluated together, as in the type I error estimates, the Pearson's correlation, Winsorized, Spearman Rank, and Kendal-Tau correlation coefficients are seem to be the most powerful coefficients. Pernet and his colleagues (2013) reported that the Pearson's correlation was the best method, estimating best the true effect sizes and showing more power as long as samples were taken from normal distribution [44]. However, the assumption of normality is generalley met and thus when it is not met, using Pearson's or Spearman's correlations can lead to serious errors. Tuğran and his colleagues (2015) reported that the Type I error rate and power of Pearson correlation coefficient were negatively affected by the distribution shapes especially for small sample sizes, which was much more pronounced for Spearman Rank and Kendal Tau correlation coefficients [4]. In conclusion, when assumptions of Pearson correlation coefficient are not satisfied, Permutation-based and Winsorized correlation coefficients seem to be better alternatives. Wilcox (1994) reported that the percentage bend correlation gave much better results in terms of controlling type I error rates comparing to the Winsorized and the Pearson's correlations when testing independence [8]. On the other hand, none of the three correlation coefficients dominates the other two in terms of power. However, he informed that no correlation coefficient was found to be dominant over the other two correlations in terms of the power of the test. Keskin and Mendeş (2021) in their simulation study they reported that especially the Pearson correlation coefficient is highly affected from outlier [42]. It was informed that when datasets contain outliers, on the other hand, probably more stable results will achieve especially when Biweight midcorrelation is preferred. They also reported that especially the Pearson, Winsorized, and Distance correlations were not affect from changes in sample sizes as long as outliers were not presence. When the results of our simulation study are compared with the literature, it can be said that our results are generally compatible with the literature, although there are some differences due to the differences in the experimental conditions of the studies.

References

[1] Carroll, J.B. (1961). The nature of the data, or how to choose a correlation coefficient. Psychometrika, 26, 347–372.

[2] Chen, P.Y., Popovich, P.M. (2002). Correlation: Parametric and Nonparametric Measures. Series:Quantitative Applications in the Social Sciences, Sage Publications, Inc., California, USA.

[3] Mendeş, M. (2019). İstatistiksel Yöntemler ve Deneme Planlanması. Birinci Baskı, Kriter Yayınları, İstanbul, 636 (in Turkish).

[4] Tuğran, E., Kocak, M., Mirtagioğlu, H., Yiğit, S., & Mendes, M. (2015). A simulation based comparison of correlation coefficients with regard to type I error rate and power. Journal of Data Analysis and Information Processing, 3 (03), 87-101.

[5] Wilcox R. R. (2012a). Introduction to Robust Estimation and Hypothesis Testing, 3rd Edn Oxford: Academic Press.

[6] Wilcox R. R. (2012b). Modern Statistics for the Social and Behavioral Sciences. Boca Raton, FL: CRC Press.

[7] Choi, J., Peters, M., & Mueller, R. O. (2010). Correlational analysis of ordinal data: from Pearson's r to Bayesian polychoric correlation. Asia Pacific Education Review, 11(4), 459-466.

[8] Wilcox R.R. (1994). The percentage bend correlation coefficient. Psychometrika 59, 601–616.

[9] Zar, J. H. (1999). Biostatistical Analysis. Fourth Edition. Simon & Schuster/A Viacom Co., New Jersey, USA.

[10] R (4.0.2). R version 4.0.2 (2020-06-22) "Taking Off Again" Copyright (C) 2020 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32/x64 (64-bit).

[11] R Studio (1.2.5033). RStudio Version 1.2.5033 – © 2009-2020 RStudio, Inc.

[12] Bishara, A.J., Hittner, J.B. (2012). Testing the Significance of a Correlation With Nonnormal Data: Comparison of Pearson, Spearman, Transformation, and Resampling Approaches. Psychological Methods, 17(3), 399-417.

[13] Bishara, A. J., Hittner, J. B. (2017). Confidence intervals for correlations when data are not normal. Behavior Research Methods, 49(1), 294–309.

[14] Fieller, E. C., Hartley, H. O., & Pearson, E. S. (1957). Tests for rank correlation coefficients. Biometrika, 44(3/4), 470–481.

[15] Zar, J.H., (2014). Spearman Rank Correlation: Overview. Wiley StatsRef: Statistics Reference Online. doi:10.1002/9781118445112.stat05964.

[16] Kendall, M., Gibbons, J.D. (1990) Rank Correlation Methods. 5th Edition, Edward Arnold, London.

[17] Knight, W.E. (1966) A Computer Method for Calculating Kendall's Tau with Ungrouped Data. Journal of the American Statistical Association, 61, 436–439.

[18] Sheskin, D. (2011). Handbook of Parametric and Nonparametric Statistical Procedure (5th ed.). Boca Raton, FL: CRC Press.

[19] Wilcox, R.R. (1993). Some Results on a Winsorized Correlation Coefficient. British Journal of Mathematical and Statistical Psychology, 46, 339-349.

[20] Wilcox, R.R. (2001). Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy. Springer, New York. http://dx.doi.org/10.1007/978-1-4757-3522-2

[21] Hoeffding, W. (1948). A Non-Parametric Test of Independence. Annals of Mathematical Statistics, 19, 546–557.

[22] Fujita, A., Sato, J. R., Demasi, M. A. A., Sogayar, M. C., Ferreira, C. E., & Miyano, S. (2009). Comparing Pearson, Spearman and Hoeffding's D measure for gene expression association analysis. Journal of Bioinformatics and Computational Biology, 7(4),663– 684.

[23] Hollander, M.,Wolfe, D. (1973), Nonparametric Statistical Methods, New York: John Wiley & Sons, Inc.

[24] Base SAS® 9.2 - Procedures Guide - DataJobs.com (Access date:).https://datajobs.com/data-science-repo/SAS-Stat-Guide-[SAS-Institute].pdf (Access date: August 5, 2021).

[25] Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35(6),2769–2794.

[26] Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. The Annals of Applied Statistics, 3(4),1236–1265. Mathematical Reviews number (MathSciNet): MR2752127.

[27] Székely, G.J., Rizzo, M.L. (2012). On the uniqueness of distance covariance. Statistics & Probability Letters, 82 (12), 2278–2282.-27-

[28] Székely, G.J., Rizzo, M.L. (2013). The distance correlation t-test of independence in high dimension. Journal of Multivariate Analysis, 117, 193-213.-28-

[29] Bhattacharjee, A. (2014). Distance Correlation Coefficient: An Application with Bayesian Approach in Clinical Data Analysis. Journal of Modern Applied Statistical Methods, 13 (1), 354-366. -29-

[30] Dueck, J., Edelmann, D., Gneiting, T.,& Richards, D. (2014). The affinely invariant distance correlation. Bernoulli, 20(4), 2305-2330. https://doi.org/10.3150/13-BEJ558 -30-

[31] Sejdinovic, D., Sriperumbudur, B., Gretton, A., & Fukumizu, K. (2013), Equivalence of distance-based and RKHS statistics in hypothesistesting. The Annals of Statistics, 41(5),2263-2291.-31-

[32] Zhong, J., DiDonato, N., &Hatcher, P.G. (2012). Independent component analysis applied to diffusion-ordered spectroscopy: separating nuclear magnetic resonance spectra of analytes in mixtures. Journal of Chemometrics, 26, 150-157.-32-

[33] Langfelder, P., Horvath, S. (2012). Fast R functions for robust correlations and hierarchical clustering. Journal of Statistical Software, 46(11). doi:10.18637/jss.v046.i11 -33-

[34] Zheng, C-H., Yuan, L., Sha, W., &Sun, Z-L. (2014). Gene differential coexpression analysis based on biweight correlation and maximum clique. BMC Bioinformatics, 15 (Suppl 15):53.-34-

[35] Lin, Y., Wen, Z.L.S., & Zheng, C.H., (2013). Biweight Midcorrelation-Based Gene Differential Coexpression Analysis and Its Application to Type II Diabetes. ICIC 2013, CCIS 375, pp. 81–87.-35-

[36] R Development Core Team (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.-36-

[37] Blomqvist, N. (1950). On a measure of dependence between two random variables. Annals of Mathematical Statistics, 21, 593–600.-37-

[38] Schmid, F., Schmidt, R. (2007). Nonparametric Inference on Multivariate Versions of Blomqvist's Beta and Related Measures of Tail Dependence. Metrika, 66(3),323-354.-38-

[39] Mosteller and Tukey (1977). Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, pp. 203-209.-39-

[40] Shoemaker and Hettmansperger (1982). Robust Estimates of and Tests for the One- and Two-Sample Scale Models, Biometrika 69, 47-54.-40-

[41] Wilcox, R.R. (1997). Introduction to Robust Estimation and Hypothesis Testing. Academic Press.-41-

[42] Keskin, E., Mendeş, M. (2021). Comparing Different Correlation Coefficients over Large Samples. IV. International Conference on Data Science and Applications (ICONDATA'21), June 4-6, 2021, TURKEY. -42-

[43] Kraemer, H.C. (2006). Correlation coefficients in medical research: from product moment correlation

to the odds ratio. Statistical Methods in Medical Research, 15(6),525-545. -43-

[44] Pernet, C.R., Wilocox, R.R., & Rousselet, G.A. (2013). Robust correlation analyses: false positive and power validation using a new open source Matlab toolbox. Frointiers in Psychology, 3, 1-18.-44-

[45] Zhou, Y., Zhang, Q., &Singh, V.P. (2016). An adaptive multilevel correlation analysis: a new algorithm and case study. Hydrological Sciences Journal–Journal Des Sciences Hydrologiques, 61(15), 2718-2728.-45-

[46] Bradley, J. C. (1978). Robustness. British Journal of Mathematical and Statistical Psychology, 31, 144-152. -46-

[47] Cochran, W. G. (1954). Some methods for strengthening the common $\chi 2$ -tests. Biometrics, 10, 417-451. -47-