ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2018-59714

Evaluation of Sample Size Effect on Spearman and Polyserial Correlation Coefficients

Spearman ve Poliserial İlişki Katsayıları Üzerine Örneklem Büyüklüğünün Etkisinin Değerlendirilmesi

Selcen YÜKSEL[a]

[a]Department of Biostatistics, Ankara Yıldırım Beyazıt University Faculty of Medicine, Ankara

Received: 10.01.2018
Received in revised form: 27.02.2018
Accepted: 09.03.2018
Available online: 05.04.2018

Correspondence:
Selcen YÜKSEL
Ankara Yıldırım Beyazıt University Faculty of Medicine, Department of Biostatistics, Ankara, TURKEY/TÜRKİYE
selcenpehlivan@gmail.com

**A different version of this study as contextual was presented at IBS-EMR 2015 Cappadocia as poster presentation with title" A Simulation Study: Correlation between an Ordinal and a Continuous Variable, Spearman or Polyserial?"*

ABSTRACT Correlation analysis, which examines whether there is a relationship between two or more variables, is a statistical method that is widely used in many areas such as in the health field. When analyzing the relationship between variables, it is important to know which correlation coefficients should be used according to variable types. If one of the two variables considered to be in a linear relationship does not show normal distribution, the linear relationship between the order of the variables is evaluated by the Spearman rho correlation coefficient. In practice, the relationship between an ordinal and a continuous variable is generally evaluated by the Spearman rho correlation coefficient, assuming that the ordinal variables do not fit the normal distribution. However, in the literature, the necessity of using the polyserial correlation coefficient between an ordinal and continuous variable is mentioned. The purpose of this study is to examine the extent to which the results obtained from the polyserial correlation coefficients are separated from the Spearman rho correlation coefficient for different sample sizes. For this purpose, the separation of polyserial correlation coefficients from the Spearman rho correlation coefficient, which is derived from 10,000 MCMC data for each of the 30 different sample sizes, has been investigated. If the sample size is less than 500, it is observed that the discrepancies are increased, and that the confidence intervals of these discrepancies become larger. It is not always possible to work with large samples in the health field. Especially in studies in the health field, if the sample size is less than 500, the correlation between an ordinal and a continuous variable must be evaluated with the polyserial correlation coefficient.

Keywords: Polyserial; Spearman; simulation; sample size

ÖZET İki ya da daha çok değişken arasında ilişki olup olmadığını inceleyen ilişki analizi, sağlık alanında olduğu gibi birçok alanda çok yaygın olarak kullanılan istatistiksel bir yöntemdir. Değişkenler arası ilişki incelenirken, değişken türlerine göre hangi ilişki katsayılarının kullanılması gerektiğinin bilinmesi önemlidir. Aralarında doğrusal ilişki olduğu düşünülen iki değişkenden biri normal dağılıma uygunluk göstermiyorsa, değişkenlerin sıraları arasındaki doğrusal ilişki Spearman rho ilişki katsayısıyla değerlendirilir. Pratikte genellikle, bir sıralı ve bir sürekli değişken arasındaki ilişki, sıralı değişkenlerin normal dağılıma uymadığı varsayımı altında, Spearman rho ilişki katsayısıyla değerlendirilir. Fakat literatürde bir sıralı bir sürekli değişken arasındaki ilişkiyi incelemede poliserial ilişki katsayısının kullanılması gerekliliğine değinilmiştir. Bu çalışmanın amacı, poliserial ilişki katsayılarından elde edilen sonuçların hangi örneklem büyüklüklerinde Spearman rho ilişki katsayısından ayrıldıklarının sorgulanmasıdır. Bu amaca yönelik, 30 farklı örneklem büyüklüğünün her biri için 10000 veri türetilmiş, poliserial ilişki katsayılarının Spearman rho ilişki katsayısından ayrılmaları incelenmiştir. Örneklem büyüklüğünün 500'ün altında olması durumunda, ayrışmaların arttığı ve bu ayrışmaların güven aralıklarının da genişlediği gözlemlenmiştir. Sağlık alanında büyük örneklemler ile çalışmak her zaman mümkün olmamaktadır. Özellikle sağlık alanında yapılan çalışmalarda, örneklem büyüklüğünün 500'ün altında olması durumunda bir sıralı ve bir sürekli arasındaki ilişki mutlaka poliserial ilişki katsayısı ile değerlendirilmelidir.

Anahtar Kelimeler: Poliserial; Spearman; benzetim; örneklem büyüklüğü

In statistics, correlation analysis helps to determine the relationship between two random variables. The correlation coefficient obtained from the correlation analysis ranges between -1 and +1. The sign and magnitude of this coefficient indicate the direction and the strength of the association, respectively. The strength of the correlation can be interpreted by using boundaries for the absolute value of the coefficient suggested by Evans in 1996.[1] According to this classification, when the value of a correlation ranges from 0.00 to 0.19, from 0.20 to 0.39, from 0.40 to 0.59, from 0.60 to 0.79, and from 0.80 to 1.00, the strength of the relationship is described as "very weak", "weak", "moderate", "strong" and "very strong", respectively.

There are many types of correlation coefficient according to type of variables.[2] It should be decided which correlation coefficient is more appropriate in which condition. Pearson's correlation coefficient, originated by Karl Pearson in 1900, is the most widely used statistical method to measure the linear relationship between two normally distributed variables.[3] When the assumptions about the distribution of two continuous variables are not normal, the commonly used correlation coefficient is the Spearman rank correlation, which is a non-parametric measure using a monotonic function.[4] This coefficient assesses the monotonic relationship between two variables (the linear correlation of the ranks of the variables) that are both continuous and ordinal. In the literature, it is stated that the Polyserial Correlation Coefficient must be used to test the correlation between an ordinal and a continuous variable. Polyserial correlation was first introduced by Olsson (1982) and compared with biserial correlation.

A search of the PubMed and Scopus databases was performed to identify original research reports dealing with simulation studies on the comparison of Spearman and polyserial coefficient-related terms *("polyserial" AND "Spearman" AND "simulation")* in the title or abstract. There was no simulation study evaluating the difference between polyserial and Spearman correlation coefficients calculated for the correlation between an ordinal and continuous variable.

We aimed to investigate the effect of sample size on the decision of whether to use the Spearman Correlation Coefficient or Polyserial Correlation Coefficient to obtain correlation between an ordinal and a continuous variable in practice. For this aim, a simulation study was carried out to test the discrepancy between these two correlation coefficients for the number of 30 different sample sizes. 10,000 MCMC iterations were performed for each sample size.

MATERIAL AND METHODS

AFOREMENTIONED CORRELATION COEFFICIENTS

Polyserial correlation measures the correlation between two continuous variables with a bivariate normal distribution, where one variable is observed directly, and the other is unobserved. Information about the unobserved variable is obtained through an observed ordinal variable that is derived from the unobserved variable by classifying its values into a finite set of discrete, ordered values.[5,6] There are three estimation methods for the polyserial correlation coefficient; MLE, the two-step estimator and the ad hoc estimator. As Olsson et al. described in their simulation study, MLE, two-step, and ad hoc estimators all have relatively small biases and can be used interchangeably. In this study, the two-step estimator was used to calculate polyserial correlation.

The two-step estimator is obtained by estimating population mean (μ) by sample mean ($\overline{X}$), population variance (σ^2) by sample variance (s^2), and inverse values of the normal distribution function evaluated at the cumulative marginal proportions of the ordinal variable are taken as estimates of the thresholds. A

conditional maximum likelihood estimate of polyserial correlation (ρ_p) is then computed, given the other parameter estimates. This procedure is termed the two-step method.[5] The estimate of ρ_p is obtained by maximizing (1) with respect to ρ_p only, by setting (2) equal to zero and solving for ρ_p.

$$l = logL = \sum_{i=1}^{N}[logp(x_i) + logp(y_i|x_i)] \quad (1)$$

$$\frac{\partial l}{\partial \rho} = \sum_{i=1}^{N}\left\{\frac{1}{p(y_i|x_i)}\frac{1}{\sqrt[2]{(1-\rho)^3}}\left[\varphi\left(\tau_j^*\right)\left(\tau_i\rho - z_i\right) - \varphi\left(\tau_{j-1}^*\right)\left(\tau_{j-1}\rho - z_i\right)\right]\right\} \quad (2)$$

In (1) and (2) X is a continuous variable, Y is an ordinal variable, x_i and y_i are values of X and Y respectively, and τ_{j-1}^* and τ_j^* are the thresholds surrounding Y where $z_i = (x_i - \mu)/\sigma$.

The Spearman's rank correlation coefficient, ρ_s, proposed by Charles Spearman in 1904, is appropriate when variables are skewed or ordinal.[7] This non-parametric correlation coefficient is equal to the Pearson correlation of ranks. **Assuming** that x_i and y_i are values of the X and Y variables, the ρ_s can be calculated by using equation (3)

$$\rho_s = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \quad (3)$$

where d_i is the difference between ranks for X and Y.

DESIGN OF SIMULATION STUDY

5-level ordinal variables from multinomial distribution and continuous variables drawn from a uniform distribution (min=1; max=10) were simulated for 30 different sample sizes (Table 1). Normality of the derived continuous variable was not considered because in practice it is enough if either variable is not normally distributed to use ρ_s. Even so skewness and kurtosis values of the simulated data for each sample size were given to clarify the simulated continuous data distribution. As seen in Figure 1, boundaries for skewness and kurtosis were acceptable. The *"polycor"* and *"moments"* packages and "sample" function were used for the simulation study in R i386.3.2.1 language.[8] 10,000 MCMC iterations were performed for each sample size.

TABLE 1. Sample sizes taken as simulation scenarios.

Sample Sizes		
10	180	550
20	200	600
30	220	650
40	240	700
60	260	750
80	280	800
100	300	850
120	400	900
140	450	950
160	500	1000

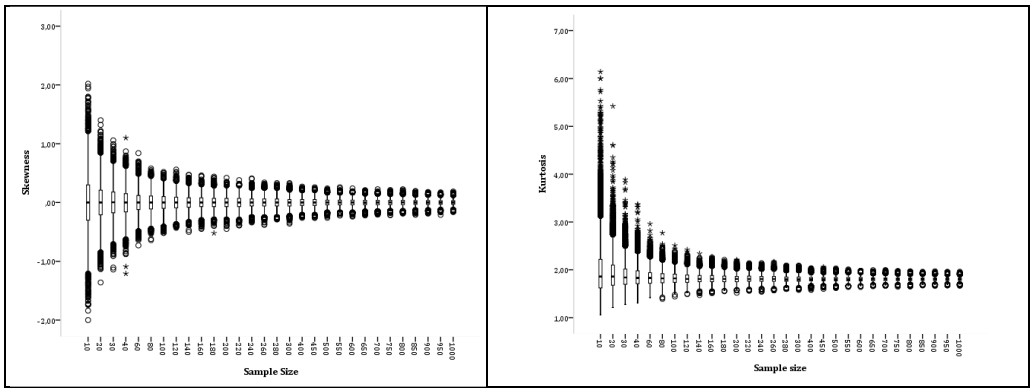

FIGURE 1. Distribution of skewness and kurtosis values of simulated continuous variables with number of 10,000 for each sample size.

Mean discrepancy between ρ_p and ρ_s (4) and 95% CI of mean discrepancy (5) were calculated to evaluate the performance of these correlation coefficients for each sample size.

$$\overline{D} = \frac{\sum_{i=1}^{10000}\left(\hat{\rho}_{s_i} - \hat{\rho}_{p_i}\right)}{10.000} \quad (4)$$

$$95\ \%CI\ of\ \overline{D}: \overline{D} \pm \left(Std.dev(\overline{D})/\sqrt{10.000} * 1.96\right) \quad (5)$$

RESULTS

Table 2 shows the results of the simulation study derived from each sample size. As seen in Table 2, the mean of discrepancy and standard deviation of discrepancy decrease as sample size increases.

TABLE 2. Simulation results for each sample size.

Sample Size	$\overline{D}$	Std. dev ($\overline{D}$)	95% CI lower bound	95% CI upper bound
10	2.91E-04	0.1057	-1.78E-03	2.36E-03
20	2.30E-04	0.0560	-8.67E-04	1.33E-03
30	2.83E-04	0.0382	-4.65E-04	1.03E-03
40	-6.18E-05	0.0298	-6.45E-04	5.22E-04
60	1.45E-04	0.0202	-2.50E-04	5.40E-04
80	5.88E-05	0.0156	-2.47E-04	3.65E-04
100	-1.91E-05	0.0128	-2.69E-04	2.31E-04
120	-1.26E-04	0.0107	-3.36E-04	8.27E-05
140	4.87E-05	0.0094	-1.36E-04	2.33E-04
160	-3.70E-06	0.0085	-1.70E-04	1.62E-04
180	3.56E-05	0.0078	-1.17E-04	1.88E-04
200	1.35E-04	0.0070	-2.56E-06	2.73E-04
220	7.94E-05	0.0065	-4.85E-05	2.07E-04
240	2.28E-05	0.0061	-9.68E-05	1.42E-04
260	-4.20E-05	0.0058	-1.55E-04	7.12E-05
280	-4.17E-05	0.0054	-1.47E-04	6.32E-05
300	1.26E-05	0.0052	-8.84E-05	1.14E-04
400	-3.26E-05	0.0042	-1.14E-04	4.91E-05
450	1.66E-05	0.0038	-5.81E-05	9.13E-05
500	3.72E-05	0.0035	-3.21E-05	1.06E-04
550	1.92E-05	0.0033	-4.54E-05	8.38E-05
600	-1.69E-05	0.0031	-7.81E-05	4.42E-05
650	5.06E-06	0.0030	-5.28E-05	6.29E-05
700	-3.06E-07	0.0028	-5.53E-05	5.47E-05
750	-1.69E-05	0.0027	-6.92E-05	3.55E-05
800	2.70E-05	0.0026	-2.42E-05	7.83E-05
850	2.07E-05	0.0025	-2.76E-05	6.91E-05
900	-1.43E-05	0.0024	-6.14E-05	3.28E-05
950	1.81E-05	0.0023	-2.72E-05	6.34E-05
1000	5.86E-06	0.0022	-3.78E-05	4.95E-05

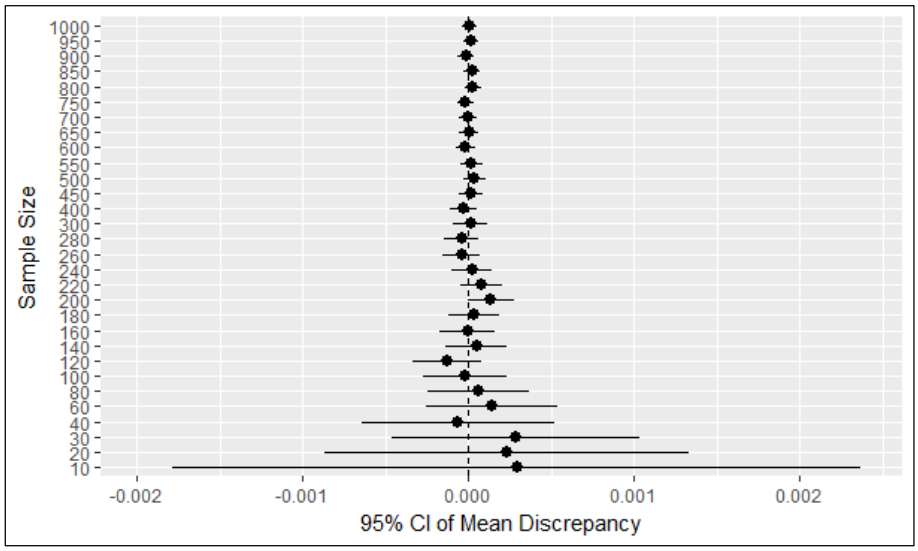

FIGURE 2. Mean discrepancy and 95% CI of mean discrepancy for each sample size

In Figure 2, the bold asterisk represents the $\overline{D}$ for each sample size. $\overline{D}$ values show how similar the ρ_s estimations are to the estimations of ρ_p. According to these values, as sample size increases, the $\overline{D}$ also approximate towards 0. The highest $\overline{D}$ was observed in the first three sample size scenarios: 10, 20 and 30. For the sample size 40, it should not be ignored that the confidence interval of $\overline{D}$ is large even if $\overline{D}$ approaches 0 much more closely. The same inferences can be made up to sample size 140. It can be said that for sample sizes greater than 500, the value of $\overline{D}$ is very close to 0. There may not be a difference between $\overline{D}$ in some sample sizes but it would be correct to interpret the confidence intervals for these sample sizes. It is obvious that as the sample sizes increase, narrow confidence intervals are obtained.

DISCUSSION

Correlation analysis is widely used in the health field as well as in many other areas. The Likert scale, which is a well-known measure of ordinal data, is commonly used to evaluate psychometric properties of a scale. The researchers can measure some latent traits of patients about a topic by using this ordinal data. When a researcher wants to evaluate the correlation between each question response obtained by a 5-point Likert type scale and total score, parametric test statistics cannot be used. In this situation, the Spearman rank correlation coefficient is a commonly used non-parametric correlation coefficient in many areas including the health field. However, the polyserial correlation coefficient is an appropriate coefficient to evaluate the association between a Likert scale/ordinal variable and a continuous variable. In this study, the decision on whether or not to use the Spearman correlation coefficient to evaluate the correlation between an ordinal variable and a continuous variable is clarified under different sample sizes. As seen in the results given above, it can be concluded that if the sample size is bigger than 500, Spearman correlation coefficients can be used instead of the polyserial correlation coefficient. If the sample size is higher than 250, there is a meaningful decrease in the discrepancy between polyserial and Spearman correlation coefficient even so confidence intervals are large until the sample size reaches 500. Hence it must be concluded that while working on small or medium sample sizes, polyserial correlation

must be preferred to calculate the relation between an ordinal variable and a continuous variable. In this study distribution of simulated continuous data did not taken as simulation condition because main idea is constructed on Spearman correlation coefficient. Results may be extended for more skewed or flat distributions. Simulated ordinal data was 5-point Likert type which is generally used for scale development or for questionnaires. For different type of ordinal data may be give different results. The correlation values were not taken as fixed simulation condition and they derivated randomly. This can be cited under the limitations of the study.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

This study is entirely author's own work and no other author contribution.

REFERENCES

1. Evans JD. Straightforward Statistics for the Behavioral Sciences. 1st ed. Pacific Grove, Calif: Brooks/Cole Publishing; 1996. p.600.
2. Öztuna D, Elhan AH, Kursun N. [Correlation coefficients in medical research: review] Turkiye Klinikleri J Med Sci 2008;28(2):160-5.
3. Ekström J. An empirical polychoric correlation coefficient. Contributions to the Theory of Measures of Association for Ordinal Variables. Ph.D. Thesis. Uppsala: Acta Universitatis Upsaliensis; 2009.
4. Mukaka MM. Statistical corner: a guide to appropriate use of correlation coefficient in medical research. Malawi Med J 2012;24(3):69-71.
5. Olsson U, Drasgow F, Dorans J. The polyserial correlation coefficient. Psychometrika 1982;47(3):337-47.
6. Drasgow F. Polychoric and polyserial correlations. Encyclopedia of Statistics. Vol. 7. New York: John Wiley; 1986. p.68-74. doi: 10.1002/0471667196.ess2014.pub2
7. Spearman C. The proof and measurement of association between two things. Am J Psychol 1904;15:72-101.
8. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. Available at: http://www.R-project.org. Accessed August 1, 2017.