Current issue list available at AnnMedRes

Annals of Medical Research

journal page: www.annalsmedres.org

Comparison via simulation of association coefficients calculated between categorical variables

Sinem Sensoy[a], Yeliz Kasko Arici[b,*]

[a] Ordu University, Institute of Natural Sciences, Ordu, Türkiye
[b] Ordu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Ordu, Türkiye

ARTICLE INFO

Keywords:
Categorical variable
Cross-table
Simulation
Ordinal data
Association coefficient

Received: Aug 18, 2023
Accepted: Sep 21, 2023
Available Online: 27.09.2023

DOI:
10.5455/annalsmedres.2023.08.206

Abstract

Aim: The aim of this study was to compare the robustness of some association coefficients used to determine the relationships between categorical variables under different experimental conditions.

Materials and Methods: A simulation study was conducted where random numbers were generated from a bivariate standard normal distribution with correlations of 0.5 and 0.9. Sample sizes were set at 30, 50, 100, 150 and 200. Random numbers were equally spaced and coded as 3×3, 4×4 and 5×5 cross-tabulations, respectively. The robustness of Pearson's, Spearman's rank, Kendall's tau-b, Kendall's tau-c, Goodman-Kruskal's gamma and Somer's d coefficients were compared under different experimental conditions consisting of combinations of specified population correlation degrees, table dimensions and sample sizes.

Results: The Goodman-Kruskal's gamma coefficient gave the closest result to the relationship levels set at the beginning of the study in all experimental conditions. However, after a certain level, it was negatively affected by the increase in table dimension and sample size. Kendall's tau-b and tau-c coefficients were furthest from the actual degree of the association. Spearman's rank correlation was more robust than Kendall's tau-b, Kendall's tau-c and Somer's d coefficients.

Conclusion: The results of the study showed that the dimension of the contingency tables and sample size were effective factors in the robustness of association coefficients for categorical variables. Therefore, researchers should consider the table dimension and sample size as well as the type of variable when selecting the association coefficient to be calculated.

Introduction

In research, it is common practice to investigate multiple variables to gain a comprehensive understanding of the event or phenomena under study. This is due to the possibility that the variables may be related to each other. To measure the strength and summarize the potential relationships between variables, coefficients called correlation or association coefficients are calculated. These coefficients reveal whether there is a relationship between the variables and whether when the value of one variable increases, the value of another tends to increase or decrease, progressively. Numerous coefficients have been developed that they are suggested for computation based on the variable type, distribution shape and sample size [1].

A categorical variable can take either nominal or ordinal form. Nominal variables contain two or more categories without any intrinsic order. In contrast, ordinal variables contain an order of magnitude or smallness from top to bottom or bottom to top [2]. The appropriate association coefficient for categorical variables depends on the number and ranking among the categories. For nominal variables, the Phi coefficient is applicable for those with two categories, while Cramer's V coefficient or Lambda (λ) coefficient is used for those with more than two categories. Concerning ordinal variables, it may use various association coefficients such as Spearman's rank, Kendall's tau-b, Kendall's tau-c, Goodman-Kruskal's gamma or Somer's d, contingent on the circumstances. It can be a problem for researchers to decide which of these association coefficients should be preferred [1].

Pearson's correlation coefficient is widely used, but its use and dependability are based on meeting its assumptions, as it is a parametric coefficient. In cases where its assumptions are not met, Spearman's rank correlation coefficient is widely used in practice. In fact, the efficacy of Spear-

*Corresponding author:
Email address: yelizkasko@yahoo.com (Yeliz Kasko Arici)

man's rank correlation coefficient, which was developed for the examination of the relationship between two ordinal variables, in each experimental condition remains unclear. For example, in non-normally distributed datasets, one study found that Spearman's rank correlation coefficient exhibited greater robustness than Pearson's [3].

In contrast, another study reported that Pearson's, Winsorize, and permutation-based correlation coefficients displayed higher robustness than Spearman's rank and Kendall's tau coefficients in diverse experimental conditions [4].

The aim of this study was to empirically compare the robustness of Pearson, Spearman's rank, Kendall's tau-b, Kendall's tau-c, Goodman-Kruskall's gamma and Somer's d coefficients under different experimental conditions using a simulation approach. The experimental conditions were designed by combinations of different correlation degrees of population, contingency table (cross-table) dimensions and sample sizes.

Materials and Method

The material of the study consisted of random numbers generated by a Monte-Carlo simulation technique from a bivariate normal distribution with a simulation program written in Fortran programming language with the support of MSDEV and subroutine BNRDF of IMSL library. The sample sizes of the two random variables with a correlation of 0.5 and 0.9 were set at as 30, 50, 100, 150 and 200, respectively.

Simulation study

A simulation program was prepared for conducting the calculations in the study. Initially, the program contained codes for generating two generated random variables from the bivariate normal distribution, with 0.5 and 0.9 correlations. The sample sizes used were 30, 50, 100, 150, and 200, respectively. The generated random numbers were equally divided and structured into cross-table dimensions of 3×3, 4×4, and 5×5. In continuation, calculations of the association coefficients to be studied were added to the program. The association coefficients were calculated under different experimental conditions. These conditions were designed by combinations of different population correlation levels, contingency table (cross-table) dimensions and sample sizes (Table 1).

The simulation program was executed 50,000 times and the coefficients were averaged. The criteria for comparison used were the level of proximity to the actual correlation level in the population.

Table 1. Experimental conditions considered for the association coefficients.

Association coefficients	Pearson, Spearman's rank, Kendall's tau-b, Kendall's tau-c, Goodman and Kruskal's gamma, Somer's d
Cross-table dimension	3×3, 4×4, 5×5
Sample size (n)	30, 50, 100, 150, 200
Association coefficients in population (ρ)	0.50, 0.90
Number of simulation	50.000

The association coefficients examined in the study Pearson's Correlation Coefficient: Pearson's correlation (rxy) coefficient is calculated to investigate the degree and direction of the relationship between the X and Y variables, both of which were obtained with the interval scale. [5]. For a bivariate dataset with n observations as shaped Pearson's correlation coefficient is calculated by Eq. (1) [4].

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} = \frac{\sum d_x d_y}{\sqrt{(\sum d_x^2)(\sum d_y^2)}} \tag{1}$$

Where

r_{xy} : Pearson's correlation coefficient

$\sum d_x d_y$: The sum of the cross-product of XY

$\sum d_x^2$: The sum of squares for variable X

$\sum d_y^2$: The sum of squares for variable Y

The set of hypotheses established for the statistical significance control of the calculated correlation coefficient is as follows:

H_0: $\rho xy = 0$

H_1: $\rho xy \neq 0$

The test statistic required to test the null hypothesis (H0) is as in Eq. (2). If the t-test statistic calculated in this way is greater than or equal to the (n-2) degrees of freedom t-table value, the H_0 hypothesis is rejected; otherwise, it is accepted [4].

$$t = \frac{r_{xy}}{\sqrt{\frac{1-r_{xy}^2}{n-2}}} \tag{2}$$

Spearman's Rank Correlation Coefficient: Spearman's rank (rs) coefficient is a non-parametric correlation that is calculated based on the ranks of the values of the X and Y variables, instead of their actual observation values. It examines the correspondence between the order of values in the X variable and that of the Y variable. This coefficient is applicable when at least one of the variables is based on ordinal scales, normal distribution is absent in at least one variable, and/or the sample size is small. It is calculated using Eq. (3) [6].

$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)} \tag{3}$$

Where

r_S: Spearman's rank coefficient

D_i: X_i-Y_iThe difference between the rank numbers of the pairs X_i and Y_j

If the sample size is larger than 10, the hypothesis test for Spearman's rank coefficient involves calculating the t-value, as in Pearson's correlation coefficient. If the resultant value is greater than or equal to the t-table value with (n-2) degrees of freedom, then the null hypothesis is rejected, and a significant relationship between the variables is considered. If the sample size is small, the "critical values in the Spearman's Rank Table," which are prepared based on the sample size, are utilised. The hypothesis control is done with the help of Eq. (4) and (5) [7].

$$S_r = \sqrt{\frac{1-r_s^2}{n-2}} \tag{4}$$

$$t = \frac{r_s}{\sqrt{\frac{1-r_s^2}{n-2}}} \tag{5}$$

Where

Sr: Standard error of correlation coeffiecient

t: t test statistics

Kendall's Tau-b (τ_b) Coefficient: Kendall's tau-b (τ_b) coefficient is more commonly used in 2×2 cross-tabulations. It adjusts for equal ranks in the dependent and independent variables. As seen in Eq. 6, the difference between the number of concordant pairs and the number of discordant pairs (P-Q) is divided by the geometric mean of the number of unequal pairs in the independent variable and the number of unequal pairs in the dependent variables. When there is statistical independence, it takes the value 0; For quadratic tables (equal numbers of columns and rows) only, it takes the value 1 or -1 if all values are located on a single diagonal [8-11]. Its statistical significance is tested by Eq. (7). If the calculated Z value exceeds a critical value, the null hypothesis is rejected, and it is concluded that there is a statistically significant relationship between the variables [5].

$$T_b = \frac{(P-Q)}{\sqrt{(N^2 - \sum_{i=1}^{I} n_{i+}^2)(N^2 - \sum_{j=1}^{J} n_{+j}^2)}} \tag{6}$$

$$Z = \frac{3T_b\sqrt{N(N-1)}}{\sum 2(2N+5)} \tag{7}$$

Where

T_b: Kendall's tau-b coefficient

P: Numbers of concordant pairs

Q: Numbers of discordant pairs

N: Number of total obserbation

${n^2}_{+j}$: The square of column sum of the jth column

${n^2}_{i+}$: The square of row sum of the ith row

Kendall's Tau-c Coefficient: One of the names "Kendall Stuart's Tau c" or "Stuart's Tauc" can also be encountered in the literature. Kendall's tau-c (τc) coefficient is a modification of Kendall's tau-b developed for large tables and specifically for non-quadratic contingency tables. Its calculation is done using Eq. (8). Its statistical significance is tested by Eq. (9). If the calculated Z value exceeds a critical value, the null hypothesis is rejected, and there is a significant relationship between the variables [5,12,13,14,20].

$$T_c = \frac{2(P-Q)}{\frac{N^2(k-1)}{k}} = \frac{2k(P-Q)}{N^2(k-1)} \tag{8}$$

$$Z = \frac{T_c}{\sqrt{\frac{2(2N+5)}{9N(N-1)}}} \sim Z_\alpha \tag{9}$$

Where

Tc: Kendall's tau-c coefficient

P: Numbers of concordant pairs

Q: Numbers of discordant pairs

N: Number of total obserbation

k: The smaller of the rows or columns

Goodman-Kruskal's Gamma Coefficient: The gamma (γ) coefficient is a symmetric measure of association that can be used on ordinal-scale variables in the contingency tables. Its values can range from -1 to 1. The gamma estimator ignores tied pairs and uses only the number of concordant and discordant pairs of observations [10,11,20]. The gamma coefficient is found as in Eq. (10) [20]. If the calculated Z value obtained from Eq. (11) exceeds a critical value, the null hypothesis is rejected, and a significant relationship between the variables [15].

$$G = \frac{P-Q}{P+Q} \tag{10}$$

$$Z = G\sqrt{\frac{P+Q}{N(1-G^2)}} \tag{11}$$

Where

G: Goodman-Kruskal's gamma coefficient

P: Numbers of concordant pairs

Q: Numbers of discordant pairs

Z: Z-test statistic

N: Number of total obserbation

Somer's d Coefficient: Somer's d coefficient is an asymmetric modification of Kendall tau-b. Unlike other coefficients used in ordinal variables, it assumes that one of the variables can be determined as a dependent variable [8,16]. d is calculated if at least two nonzero cell frequencies exist in the row and column. A d-value has the range -1 ≤ d ≤ 1. Although not valid in 2×2 tables, when d is 0, it indicates that the two variables are independent of each other [17].

If X is the independent variable and Y is the dependent variable, the highest likelihood estimator of the coefficient is calculated with Eq. (12). Similarly, if Y is the independent variable and X is the dependent variable, the highest likelihood estimator of the coefficient is calculated with Eq. (13). When both variables are assumed dependent or independent, the symmetric version of Somers' d is calculated by Eq. (14) [18].

$$\hat{d}_{YX} = \frac{2(P-Q)}{N^2 - \sum_{i=1}^{i} n_{i+}^2} \tag{12}$$

$$\hat{d}_{XY} = \frac{2(P-Q)}{N^2 - \sum_{j=1}^{j} n_{j+}^2} \tag{13}$$

$$\hat{d}_{YX} = \frac{4(P-Q)}{2N^2 - \sum_{i=1}^{i} n_{i+}^2 - \sum_{j=1}^{j} n_{+j}^2} \tag{14}$$

Where

P: Numbers of discordant pairs

Q: Numbers of concordant and discordant pairs

N: Number of total obserbation

n_{i+}^2: The square of column sum of the i$^{\text{th}}$ row

n_{+j}^2: The square of column sum of the j$^{\text{th}}$ column

The significance test of Somer's d coefficient is made from Eq. (15) and Eq. (16). If the calculated Z value exceeds a critical value, the null hypothesis is rejected, indicating a statistically significant correlation between the variables [5].

$$Z = \frac{\hat{d}_{YX}}{\sqrt{\frac{4(c^2-1)(r+1)}{9Nc^2(r-1)}}} \tag{15}$$

$$Z = \frac{\hat{d}_{XY}}{\sqrt{\frac{4(r^2-1)(c+1)}{9Nr^2(c-1)}}} \tag{16}$$

Where

c: Number of rows

r: Number of columns

N: Number of total obserbation

Z: Z test statistic

Results

The association coefficients calculated in the 3×3 cross-table are given in Table 2. Among the association coefficients, the gamma coefficient was the closest to the actual degree of population in all sample sizes. As the sample size increased, it was seen that the results were slightly higher than the actual degree. Other coefficients gave similar results and were below the degree of actual association in all sample sizes.

In the 3x3 cross-table, after the gamma coefficient, Pearson's correlation and Spearman's rank correlation coefficients came closest to the actual degree of relationship but were slightly below. For ρ =0.5, Kendall's tau-b, Kendall's tau-c, and Somer's d coefficients were similar to each other lower than the expected actual degree. In contrast, for $\rho = 0.9$, Kendall's tau-b and Somer's d coefficients were similar, but Kendall's tau-c coefficient was the furthest from the actual degree.

When the findings obtained for the 4×4 cross-table in Table 3 are examined, it was seen that all the coefficients approach the actual degree of the association of populations more than 3×3 cross-table. The gamma coefficient reached the actual association level of the population when n=100 for $\rho = 0.5$ and n=50 for $\rho = 0.9$. Even though it deviated from the actual values as the change in sample size, it was seen that it gave closer results than the other coefficients. The coefficient furthest from the actual association levels was Kendall's tau-c even if the sample size improves. The results of Kendall's tau-b and Somer's d coefficients were approximately the same, but below the population levels.

In the 4x4 cross-table, Pearson's and Spearman's coefficients were the robust correlation coefficients after the gamma coefficient. Kendall's tau-c was the first coeffi-

Table 2. The association coefficients calculated for the 3×3 cross-table.

		n				
		30	50	100	150	200
ρ=0.5	Pearson	0.406	0.400	0.398	0.396	0.396
	Spearman rank	0.405	0.400	0.398	0.396	0.395
	Kendall tau-b	0.364	0.367	0.370	0.371	0.369
	Kendall tau-c	0.365	0.367	0.370	0.371	0.369
	Gamma	0.481	0.511	0.550	0.569	0.580
	Somer d	0.364	0.367	0.370	0.371	0.369
ρ=0.9	Pearson	0.736	0.735	0.732	0.731	0.729
	Spearman rank	0.736	0.735	0.732	0.730	0.728
	Kendall tau-b	0.715	0.714	0.712	0.709	0.710
	Kendall tau-c	0.621	0.619	0.589	0.551	0.545
	Gamma	0.915	0.920	0.929	0.932	0.933
	Somer d	0.715	0.714	0.712	0.709	0.710

Table 3. The association coefficients calculated for the 4×4 cross-table.

		n				
		30	50	100	150	200
ρ=0.5	Pearson	0.431	0.432	0.434	0.434	0.431
	Spearman rank	0.430	0.431	0.432	0.433	0.430
	Kendall tau-b	0.361	0.378	0.380	0.382	0.386
	Kendall tau-c	0.342	0.340	0.339	0.330	0.327
	Gamma	0.464	0.470	0.500	0.519	0.520
	Somer d	0.361	0.378	0.380	0.382	0.386
ρ=0.9	Pearson	0.802	0.800	0.798	0.793	0.797
	Spearman rank	0.802	0.800	0.799	0.793	0.797
	Kendall tau-b	0.730	0.732	0.735	0.735	0.736
	Kendall tau-c	0.661	0.655	0.635	0.624	0.620
	Gamma	0.899	0.900	0.910	0.915	0.920
	Somer d	0.730	0.732	0.735	0.735	0.736

Table 4. The association coefficients calculated for the 5×5 cross-table.

		n				
		30	50	100	150	200
ρ=0.5	Pearson	0.450	0.452	0.451	0.440	0.438
	Spearman rank	0.451	0.452	0.450	0.440	0.438
	Kendall tau-b	0.480	0.487	0.490	0.485	0.484
	Kendall tau-c	0.350	0.349	0.348	0.340	0.331
	Gamma	0.485	0.500	0.519	0.521	0.530
	Somer d	0.480	0.487	0.490	0.485	0.484
ρ=0.9	Pearson	0.835	0.830	0.820	0.805	0.800
	Spearman rank	0.834	0.830	0.820	0.805	0.801
	Kendall tau-b	0.770	0.765	0.759	0.749	0.748
	Kendall tau-c	0.710	0.700	0.689	0.670	0.651
	Gamma	0.920	0.923	0.930	0.940	0.940
	Somer d	0.770	0.765	0.760	0.749	0.748

cient, that was the furthest from the degree of association determined in the population, even though the sample size increased, while Kendall's tau-b was the second coefficient. It was observed that Somer's d and Kendall's tau-b correlation coefficients gave similar results to the table dimension 3×3 cross-table.

Table 4 shows the association coefficients calculated for the 5×5 cross-table. It seems that the increase in table dimension has a positive effect on approaching the actual degree of association of the population. The gamma reached the actual degree of association of the population when $\rho = 0.5$ and the sample size was 50. It also gave the closest result for $\rho = 0.9$, although not exactly. However, it gives results above the actual degree of association depending on the increase in sample size. This increase also caused Pearson's, Spearman's rank and Kendall's tau-c coefficients to diverge from the actual value and to obtain lower values.

Kendall's tau-b and Somer's d coefficients calculated to 5×5 cross-table showed similar results for both $\rho = 0.5$ and $\rho = 0.9$ at all sample sizes. After the gamma coefficient, the coefficients closest to the actual correlation level were Kendall's tau-b and Somer's rank coefficients for $\rho = 0.5$, Pearson's and Spearman's rank coefficients for $\rho = 0.9$. Pearson's and Spearman's rank coefficients tended to diverge from the actual degree of association with increasing sample size. Kendall's tau-c was the correlation coefficient that deviated the most from the actual degree of correlation in all sample sizes even if the sample size increased.

Discussion

In one study, it was emphasized that Kendall's tau-b, Kendall's tau-c, and Somer's d coefficients always give results below the determined correlation levels, even as the table size increases. In contrast, the gamma coefficient, it was emphasized that if the table dimension is small, it is good for a small sample size. The gamma coefficient approximates the actual degree of association for square tables. When the table dimension is large enough, Pearson's correlation and Spearman's rank coefficient are closer to the actual degree of association regardless of sample size, however, coefficients other than the gamma are always below the actual relation level [19]. In another study, it was emphasized that as the size of the table increases, Spearman's rank coefficient gives results close to the actual correlation level and it is appropriate to use this coefficient because it gives more consistent results in measuring the relationship between variables [20].

Conclusion

The present tried to explain in which situation the researchers should calculate which association coefficients, and the robustness of some association coefficients under different experimental conditions were compared with the simulation approach. The results showed that Goodman Kruskal's gamma coefficient gave the closest result to the actual relationship levels found in the population in all experimental conditions. However, after a certain level, it was negatively affected by the increase in table dimension and sample size. Kendall's tau-b and tau-c coefficients were furthest from the actual levels of association. Spearman's rank correlation was more robust than Kendall's tau-b, Kendall's tau-c and Somer's d coefficients. Spearman's rank and Pearson's correlation coefficients were always almost equal and more robust than Kendall's tau-b, Kendall's tau-c, and Somer's d coefficients. The findings of the study were generalized, it was decided that both table dimension and sample size were effective factors on the robustness of the association coefficients. For this reason, it is recommended that researchers consider the table dimension and sample size as well as the variable type when choosing the association coefficient to calculate.

Disclosure

This study was produced from the master's thesis conducted by the first author under the supervision of the second author at Ordu University.

Ethical approval

This is a study that does not require ethics committee approval.

References

1. Şensoy S. Comparison via simulation of the coefficients of association between categorical variables. Master's thesis, Ordu University, Ordu, Turkiye, 2020.
2. Sencan H. Scientific research in the social and behavioral sciences. Seckin Publishing, Ankara, Turkiye, 2007;170p.
3. Fowler R. Power and robustness in product-moment correlation. Applied Psychological Measurement 1987;11(4), 419-428.
4. Tuğran E. Comparison of some correlation coefficients in terms of Type I error rate and test power. Master's thesis, Çanakkale Onsekiz Mart University, Çanakkale, Turkiye, 2015.
5. Sheskin DJ. Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition (5th ed.). Chapman and Hall/CRC, 2011. https://doi.org/10.1201/9780429186196.
6. Hamarat B. SPSS and Minitab Applied Statistics. Çanakkale: Paradigma Akademi Publishing, 2017.
7. Mendeş M. Statistics and Research Methods for Applied Sciences. Kriter Publishing, İstanbul, 2012;629 p.
8. Öztuna D, Elhan A, Kurşun N. Correlation coefficients in medical research: review. Turkiye Klinikleri J Med Sci 2008; 28(2): 160–165.
9. Hinton PR, Brownlow C, McMurray I & Cozens B. SPSS explained. London, Newyork, 2004; 316 pp.
10. Karagöz Y. Relationship coefficients in the cross tabulation menu in SPSS 18. Detay Publishing, Ankara, 2010; 122 p.
11. Dikmen AC. Investigation of Ordinal Scale Relationship Coefficients and the Satisfaction of Students in Private Education Institutions Providing KPSS Education-Sivas Province Example. Master's thesis, Cumhuriyet University, Sivas, Turkiye, 2016.
12. De Muth JE. Basic statistics and pharmaceutical statistical aplications. Second Edition. London, Newyork, 2006; 721 p.
13. Stuart A. The estimation and comparison of association in contingency tables. Biometrika, 1953; 40:105-110.
14. Terzi Y. Statistical Data Analysis with SPSS (Lecture Notes), Ondokuz Mayıs University, Samsun, https://avys.omu.edu.tr/storage/app/public/yukselt/64430/SOS 2.pdf [Acess Date: 12/05/2023].
15. İsrael D. Data analysis in business research a step-by-step nonparametric approach. Sage Publications Inc, India, 2008; 281 p.
16. Siegel S & Castellan NJ. Nonparametric statistics for the behavioral sciences. 2nd ed. New York: McGraw-Hill, 1988;399 p.
17. Goodman LA & Kruskal WH. Measures of association for cross classifications, IV: simplification of asymptotic variances. Journal of the American Statistical Association, 1972;67:410-420.
18. Somers RH. A new asymmetric of association for ordinal variables. American Sociological Review, 1962;27:804.

19. Göktaş A. & İşçi Ö. A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. Metodoloski zvezki, 2011;8(1):17-37.
20. Altaş D, Kaspar EÇ & Ergüt Ö. Comparison of Correlation Coefficients: A Simulation Study. Papers on Social Science. 2012;5: 1-10.