Pearson's or spearman's correlation coefficient – which one to use?

Pearsonov ili Spearmanov koeficijent korelacije – koji koristiti?

Rebekić, A., Lončarić, Z., Petrović, S., Marić, S.

Poljoprivreda/Agriculture

ISSN: 1848-8080 (Online)
ISSN: 1330-7142 (Print)

DOI: http://dx.doi.org/10.18047/poljo.21.2.8

Poljoprivredni fakultet u Osijeku, Poljoprivredni institut Osijek

Faculty of Agriculture in Osijek, Agricultural Institute Osijek

ISSN 1330-7142
UDK: 631.523:633.11
DOI: 10.18047/poljo.21.2.8

PEARSON'S OR SPEARMAN'S CORRELATION COEFFICIENT – WHICH ONE TO USE?

Rebekić, A., Lončarić, Z., Petrović, S., Marić, S.

Original scientific paper
Izvorni znanstveni članak

SUMMARY

Most commonly used correlation coefficients are Pearson's product moment correlation coefficient and Spearman's rank correlation coefficient. The aim of this paper is to compare a Pearson's and Spearman's coefficient of correlation on the same data set. The winter wheat grain cadmium (Cd) concentration was correlated to grain zinc (Zn) concentration, plant height, plant weight, number of spikelets per spike and 1000 kernel weight. Data were collected from the experiment carried out in semi controlled conditions, where genotypic specificity of winter wheat varieties was tested on the grain Cd and Zn accumulation on uncontaminated and Cd contaminated soil. Results showed that selection of most convenient correlation coefficient mostly depends on the type of variables, presence of outliers normality and linearity of relationship.

Key-words: linear relationship, outliers, log 10 transformation, frequency distribution, winter wheat

INTRODUCTION

Examination of relationship between variables is quite often in agronomic research, so it is important for agronomists to understand and objectively interpret results of correlation analysis. In general, Pearson's product-moment correlation coefficient (r) and Spearman's rank correlation coefficient (r_s) are the most frequently used correlation coefficients (Udovičić et al., 2007).

In the 1896, Karl Pearson was the first one who described the coefficient of correlation (Hauke and Kossowski, 2011). Besides that, in his work he acknowledged Francis Galton's concept of correlation and Auguste Bravais's contribution (Denis, 2001) in developing mathematical theory of correlation. Pearson called this method „product-moments" method (or the Galton function for the coefficient of correlation r) and later this method was named after him. The Pearson's product-moment correlation coefficient is the strength measure of the *linear association* between variables (Hauke and Kossowski, 2011). For that reason the r coefficient of correlation is employed for variables on an interval or ratio scale (numerical data) that are in linear relation where each variable is normally distributed. Sometimes, the variables may be connected without being in linear relation and then the Pearson's correlation coefficient shouldn't be calculated (Udovičić et al., 2007). In such cases and when assumption of the bivariate normal distribution is not tenable Spearman's correlation coefficient (r_s) should be used (Artusi et al., 2002). Spearman's rank correlation coefficient, named after Charles Spearman, is a non-parametric measure of relation between variables, using ranks to calculate the correlation. Sometimes, Spearman's correlation is defined as Pearson's correlation coefficient between rank variables. While Pearson's correlation describes how well relationship between variables can be described using linear function, Spearman's correlation assesses how well the relationship between two variables can be described using a monotonic function. Linear and monotonic functions are shown in Figure 1.

Assist. Prof. Andrijana Rebekić (arebekic@pfos.hr), Prof. Dr. Zdenko Lončarić, Assist. Prof. Sonja Petrović, Prof. Dr. Sonja Marić – Josip Juraj Strossmayer University of Osijek, Faculty of Agriculture in Osijek, Kralja Petra Svačića 1d, 31000 Osijek, Croatia

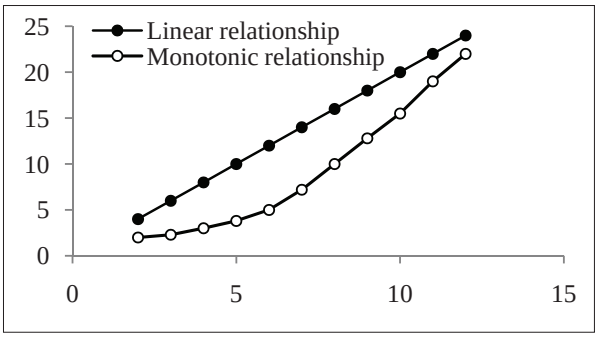

Figure 1. Linear and monotonic function
Slika 1. Linearna i monotona funkcija

Coefficients of correlation (r and $r_{s)}$ are measures of statistical correlation. Their numerical value is indicator of the strength and direction of relationship between the variables. The values of r and r_s can vary from -1.00 to 1.00. In general, $r > 0$ indicates positive relationship, $r < 0$ negative relationship while $r = 0$ indicates no relationship (or that the variables are independent and not related). According to Roemer-Orphal scale correlation coefficients are divided into classes according to their numerical value, where each class represents different strength of relationship between the variables. For example, $r = 0.78$ indicating on very strong correlation of positive direction between the variables. At the same time the strength of the correlation is not dependent on the direction or the sign. A positive correlation coefficient indicates that an increase in the first variable would correspond to an increase in the second variable while negative correlation indicates an inverse relationship whereas one variable increases the second variable decreases (Taylor, 1990). Although $r = 0.78$ indicates very strong correlation, respective relation doesn't have to be statistically significant. Statistical significance of relationship doesn't depend on the strength of correlation but mostly on the sample size and variability of the examined trait (Eđed et al., 2009). Most commonly reported correlation coefficient in scientific papers from agronomy field is Pearson's r coefficient. In many publications Pearson's correlation coefficient is reported for variables that are, by its nature or by high number of outliers, non-normally distributed. Based on this, one can put questions: is it a coincidence that majority of examined relationships between the variables are linear and/or does author's reports only significant linear relationships neglect possibility of significant non-linear relationship between variables in cases where Pearson's r coefficient is non-significant? For that reason the aim of this paper was to explain the practical connotation and limitations of Pearson's product-moment correlation coefficient (r) and Spearman's rank correlation coefficient (r_s) in examination of relationship between different types of variables.

MATERIAL AND METHODS

Details on setting up and conducting experiment are published in Eđed et al. (2010). Six variables were chosen (grain Cd concentration (mg kg^{-1}), grain Zn concentration (mg kg^{-1}), plant height (cm), plant weight (g), number of spikelets per spike and 1000 kernel weight (g)) to test differences between Pearson's r and Spearman's r_s correlation coefficient on Cd contaminated and uncontaminated soil. Grain Cd concentration was correlated to all other variables. Statistical analyses were done using SAS 9.3. Software for Windows, Copyright © 2012 by SAS Institute Inc., Cary, NC, USA, SAS Enterprise Guide 5.1 and SAS JMP® 9.0.2. Distribution of variables was tested using Shapiro-Wilks test and non-normally distributed variables were log 10 transformed to assess normality. Where it was necessary outliers were removed from data set.

RESULTS AND DISSCUSION

Grain Cd concentration was measured on interval scale (mg kg^{-1}) and according to Shapiro-Wilks test of normality grain Cd concentration marginally violates the assumption of normality ($p = 0.042$). According to Shapiro-Wilks test ($p = 0.059$) the assumption of normality cannot be rejected for the log 10 transform data of grain Cd concentration although only "weak normality" was achieved. The Q-Q plot diagrams of the examined variables on uncontaminated soil are shown in Figure 2.

On uncontaminated soil, for the three variables (grain Zn concentration, plant weight and number of spikelets per spike) out of five tested variables the assumption of normality cannot be rejected according to Shapiro-Wilks test ($p = 0.667$; $p = 0.072$ and $p = 0.821$ respectively). Variables grain Zn concentration (Figure 2 (B)) and number of spikelets per spike (Figure 2 (E)) had one outlier observation, but since normality of distribution was not violated due to its presence that single observation was not excluded from analysis. Another reason for leaving those outliers in the analysis was that such distinct value, compared to other values in the sample, could be indicator of high variability of examined traits in population of winter wheat genotypes.

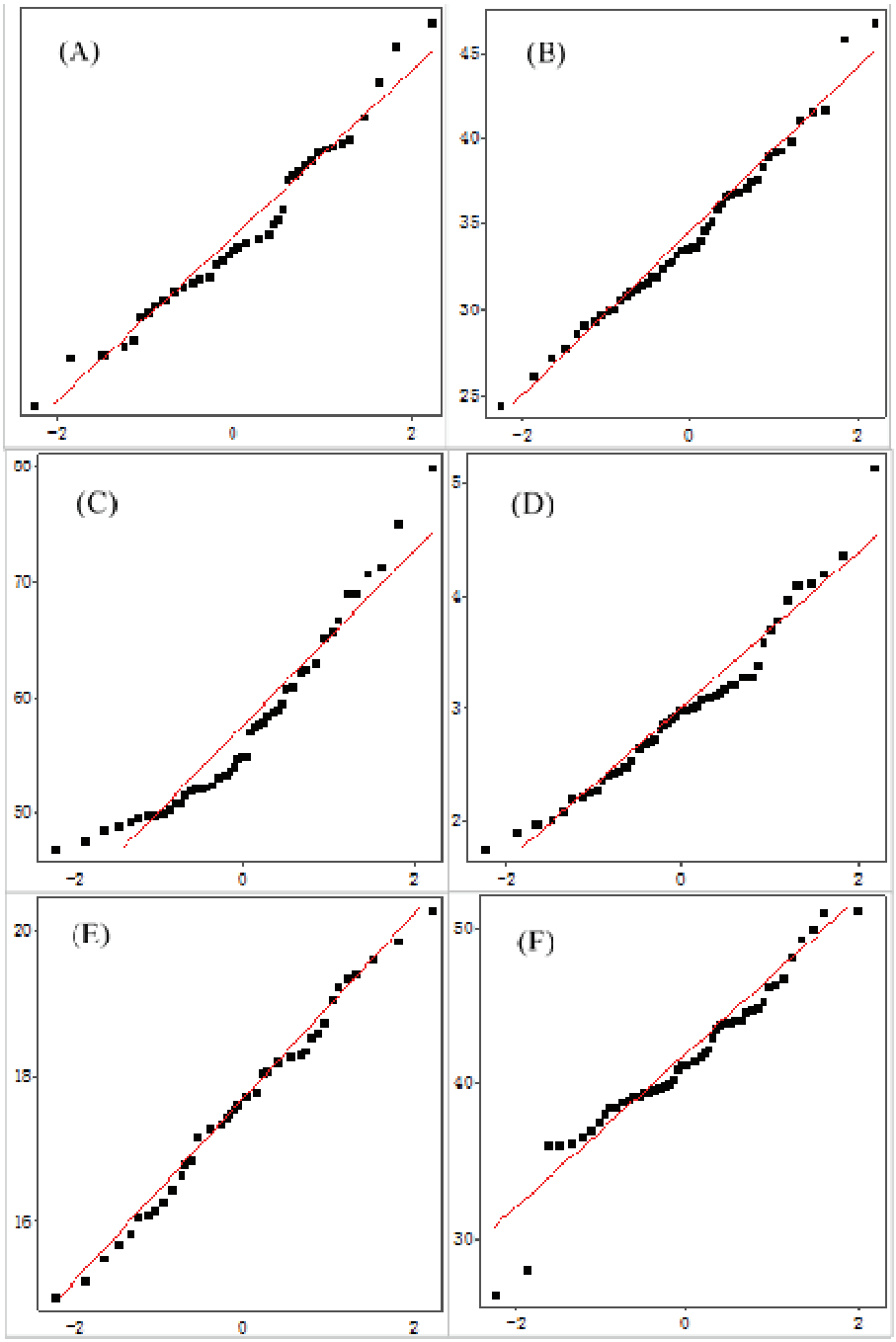

Figure 2. The Q-Q plots of (A) grain Cd concentration (mg kg^{-1}), (B) grain Zn concentration (mg kg^{-1}), (C) plant height (cm), (D) plant weight (g), (E) number of spikelets per spike and (F) 1000 kernel weight (g) for 51 winter wheat genotype on uncontaminated soil

Slika 2. Q-Q dijagrami za (A) koncentraciju Cd u zrnu (mg kg^{-1}), (B) koncentraciju Zn u zrnu (mg kg^{-1}), (C) visinu biljke (cm), (D) masu biljke (g), (E)broj klasića po klasu (F) masu 1000 zrna (g) za 51 genotip ozime pšenice na nekontaminiranome tlu

Pearson's and Spearman's correlation coefficient between grain Cd and grain Zn concentration, plant weight and number of spikelets per spike were very similar to correlation coefficients obtained from correlation analysis of log 10 grain Cd concentration and above the mentioned variables (Table 1).

According to these examples, grain Cd concentration marginally violates the assumption of normality and variables, that are correlated to that variable, follow a normal distribution. Log 10 transformation of grain Cd concentration variable, in order to asses' normality, didn't have influence on strength, direction or significance of r and r_s correlation coefficients.

Table 1. Pearson's r and Spearman's r_s correlation coefficient for grain Cd concentration (mg kg^{-1}) and log 10 grain Cd concentration and grain Zn concentration (mg kg^{-1}), plant weight (g) and number of spikelets per spike, plant height (cm) and 1000 kernel weight on uncontaminated soil (N=51)

Tablica 1. Pearson's r and Spearman's r_s koeficijent korelacije za koncentraciju Cd u zrnu (mg kg^{-1}) i koncentraciju Zn u zrnu (mg kg^{-1}), visinu biljke (cm),) masu biljke (g), broj klasića po klasu i masa 1000 zrna (g) na nekontaminiranome tlu (N=51)

	Cd concentration		Log 10 Cd concentration	
	Pearson's r	Spearman's r_s	Pearson's r	Spearman's r_s
Zn concentration	0.111 *(p =0.437)*	0.111 *(p = 0.432)*	0.112 *(p = 0.435)*	0.113 *(0.432)*
Plant weight	-0.564 *(p < 0.001)*	-0.627 *(p < 0.001)*	-0.563 *(p < 0.001)*	-0.628 *(p < 0.001)*
Number of spikelets per spike	-0.074 *(p = 0.605)*	-0.075 *(p = 0.599)*	-0.074 *(p = 0.605)*	-0.075 *(p = 0.599)*
Plant height	-0.222 *(p = 0.108)*	-0.278 *(p = 0.040)*	-0.255 *(p = 0.113)*	-0.279 *(p = 0.047)*
1000 kernel weight	0.054 *(p = 0.701)*	-0.011 *(p = 0.935)*	0,053 (p = 0,710)	-0.012 *(p = 0.935)*
Log10 plant height	-0.255 *(p = 0.084)*	-0.279 *(p = 0.047)*	-0.241 *(p = 0.087)*	-0.279 *(p = 0.047)*
Log 10 1000 kernel weight	0,058 (p = 0,688)	-0.012 *(p = 0.935)*	0.055 *(p = 0.697)*	-0.016 (p = 0.875)

On uncontaminated soil, the assumption of normality for plant height (cm) and 1000 kernel weight (g) according to Shapiro-Wilk's test has to be rejected (p = 0.002 and p = 0.025 respectively). Both variables had outliers (Figure 2) in original data set left in analysis due to their biological significance. Some genotypes included in experiment have significantly higher values compared to other genotypes included in the experiment and it was important to include these extreme values in the analysis to emphasize genotypic variability of the examined trait. The second reason for their inclusion in analysis is that in both variables, normal frequency distribution was violated, even after outliers were excluded from the analysis. After log 10 transformation of data variables, plant height and 1000 kernel weight were still non-normally distributed according to Shapiro-Wilk (p = 0.012 and p = 0.0008 respectively), and variable 1000 kernel weight had two outliers. Regarding non-normal distribution of the variable plant height, r_s coefficient would be more appropriate measure of relationship (Artusi et al., 2002), than r correlation coefficient. The r and r_s correlation coefficients for grain Cd concentration and plant height were of similar strength and direction of relationship but of different significance (Table 1). The r correlation coefficient obtained non-significant while r_s obtained significant correlation between grain Cd and plant height (in original data set). In this case it would be difficult to decide, which coefficient to report. After log 10 data transformation, correlation coefficients and significances remained almost the same (Table 1). Compared to log 10 data transformation, exclusion of outliers from the analysis showed more pronounced effect on both r and r_s coefficients (Figure 3) in correlation between grain Cd and plant height on uncontaminated soil. After exclusion of outliers r and r_s correlation coefficient between grain Cd concentration and plant height on the uncontaminated soil were of similar strength, direction and significance (r = -0.307, p = 0.033 and r_s = -0.319, p = 0.0236; N=50).

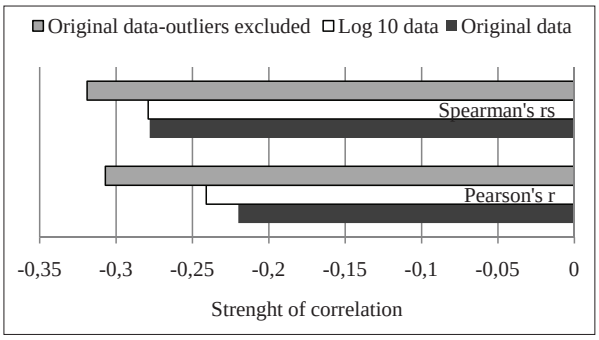

Figure 3. Pearson's r and Spearman's r_s correlation coefficients for grain Cd concentration (mg kg^{-1}) and plant height (cm) on uncontaminated soil (N = 50)

Slika 3. Pearson's r i Spearman's r_s koeficijent korelacije za koncentraciju Cd u zrnu (mg kg^{-1}) i visinu biljke na nekontaminiranome tlu (N=50)

Possible reason for differences in significance of r and r_s for grain Cd concentration and plant height on uncontaminated soil could be a sample size (Eđed et al., 2009). Changes of p value in increasing sample sizes are shown in figure 4. Increment in sample size influences only significance, not strength or direction of the relationship. In other words, with substantially large sample, every relationship can be statistically significant.

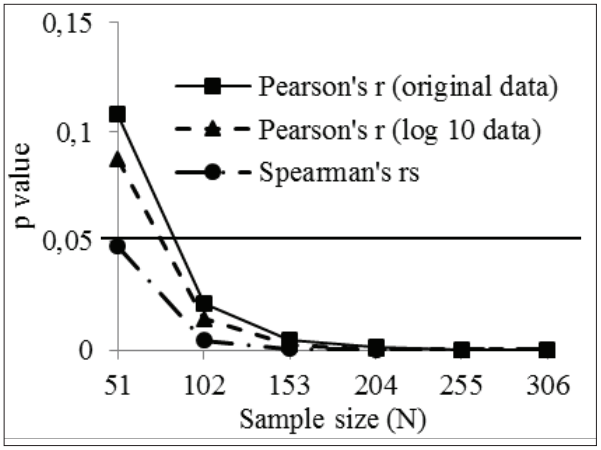

Figure 4. Statistical significance (p value) of Pearson's r and Spearman's r_s correlation coefficients between grain Cd concentration (mg kg^{-1}) and plant height (cm) in samples of different sizes

Slika 4. Statistička značajnost (p vrijednost) Pearsonovog (r) i Spearmanovog (r_s) koeficijenta korelacije za koncentraciju Cd u zrnu (mg kg^{-1}) i visinu biljke (cm) u uzorcima različite veličine

Beside that, in the interpretation of correlation between variables it is important to bear on mind that statistical significance doesn't imply biological significance and correlation doesn't mean or assume causation. In fact, coefficient of determination is better measure for determination of biological significance and according to Congelosi et al. (1983) is more meaningful than coefficient of correlation. The coefficient of determination (r^2) is defined as the percent of the variation in the values of the dependent variable that can be explained by variations in the value of the independent variable (Taylor, 1990). To clarify, Spearman's correlation coefficient (Table 1) between grain Cd concentration and plant height on the uncontaminated soil was statistically significant ($p = 0.04$) and the strength of relationship, according to Roeamer-Orphal scale, was very weak $r_s = -0.278$. Accordingly the coefficient of determination is $r^2 = 0.077$. Although the relationship is statistically significant, only 7.7% of total variations in plant height can be explained or accounted for by variation in grain Cd concentration. Undoubtedly coefficient of determination is more conservative measure of relationship between the two variables and is preferred by many statisticians, but is seldom reported (Taylor, 1990).

On Cd contaminated soil, variable grain Cd concentration followed normal frequency distribution (Figure 5 (A)) according to Shapiro-Wilk ($p = 0.794$). Among other variables on Cd contaminated soil only for the variable number of spikelets per spike (Figure 5 (E)) assumption of normally was not rejected according to Shapiro-Wilk's test ($p = 0.487$). Variable number of spikelets per spike had two outliers, one on the left and one on the right side of distribution. These outliers stretched distribution in different directions, without notable effect on the shape of curve. After log 10 transformation, according to Shapiro-Wilks test for variables grain Zn concentration ($p = 0.086$), plant height ($p = 0.247$), plant weight ($p = 0.262$) and 1000 kernel weight ($p = 0.057$) assumption of normality was not rejected, even if achieved normality was very weak for grain Zn concentration and 1000 kernel weight.

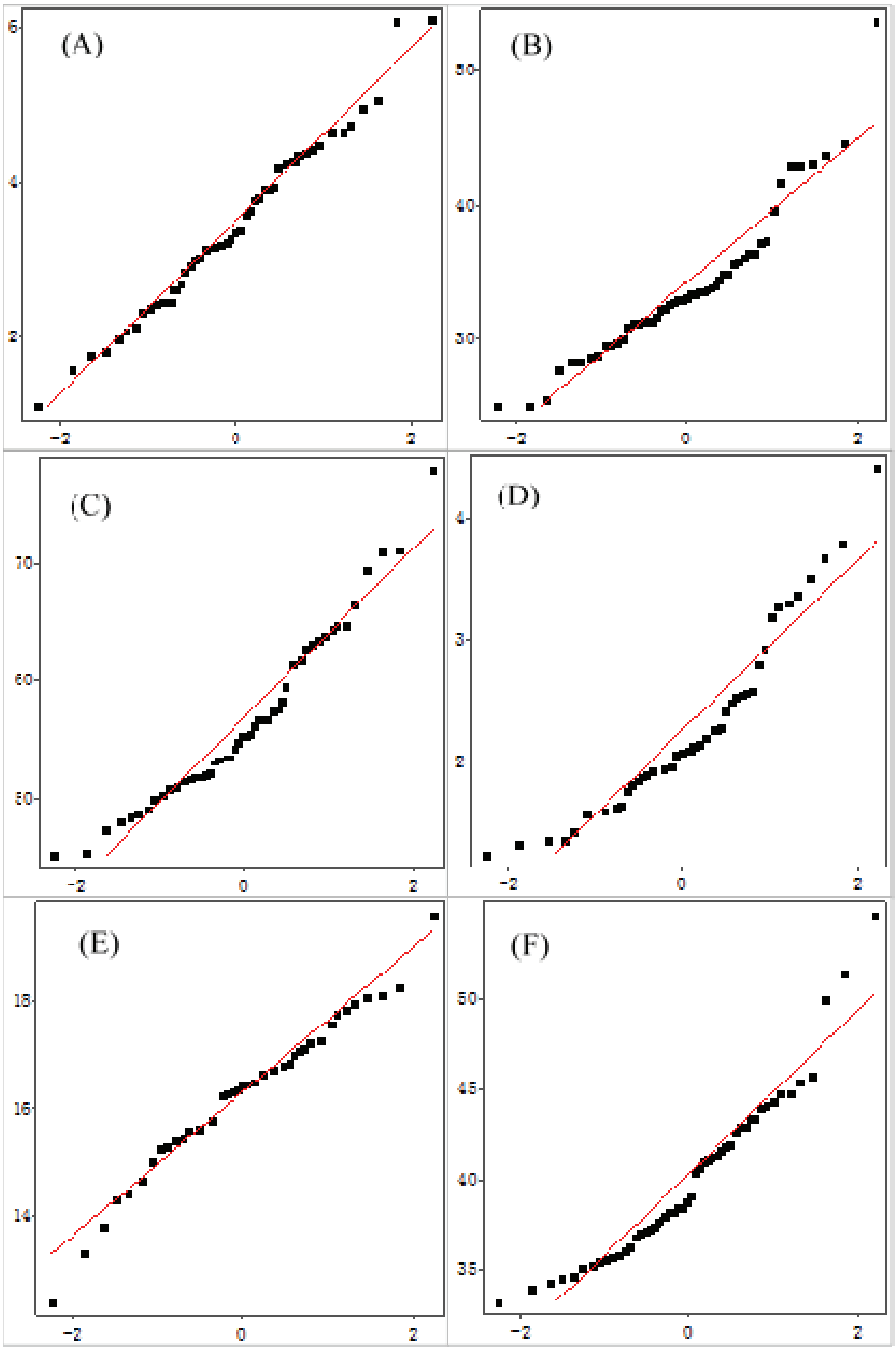

Figure 5. The Q-Q plots of (A) grain Cd concentration (mg kg^{-1}), (B) grain Zn concentration (mg kg^{-1}), (C) plant height (cm), (D) plant weight (g), (E) number of spikelets per spike and (F) 1000 kernel weight (g) for 51 winter wheat genotype on Cd contaminated soil

Slika 5. Q-Q dijagram za (A) koncentracija Cd u zrnu (mg kg^{-1}), (B) koncentracija Zn u zrnu (mg kg^{-1}), (C) visina biljke (cm), (D) masa biljke (g), (E)broj klasića po klasu (F) masa 1000 zrna (g) za 51 genotip ozime pšenice na Cd kontaminiranome tlu

On Cd contaminated soil, the most interesting results are obtained for correlation between grain Cd and Zn (Table 2). Both (r and r_s correlation coefficients) showed, according to Roemer-Orphal scale, very weak positive relationship, but r correlation coefficient showed that relationship is not significant, while r_s showed significant relationship between grain Cd and Zn concentrations on the original and log 10 transformed data. In this case, Spearman's r_s is higher than Pearson's r correlation coefficient so these two variables are in nonlinear relation, so better choice in reports would be Spearman's rank correlation coefficient.

Table 2. Pearson's r and Spearman's r_s correlation coefficient for grain Cd concentration (mg kg^{-1}) and log 10 grain Cd concentration and grain Zn concentration (mg kg^{-1}), plant weight (g) and number of spikelets per spike, plant height (cm) and 1000 kernel weight on Cd contaminated soil (N=51)

Tablica 2. Pearson's r i Spearman's r_s koeficijent korelacije za koncentraciju Cd u zrnu (mg kg^{-1}) i koncentraciju Zn u zrnu (mg kg^{-1}), visinu biljke (cm),) masu biljke (g), broj klasića po klasu i masa 1000 zrna (g) na Cd kontaminiranome tlu (N=51)

	Cd concentration	
	Pearson's r	Spearman's r_s
Zn concentration	0.208; $p = 0.142$	0.282; $p = 0.044$
Plant weight	-0.548; $p < 0.0001$	-0.457; $p = 0.0007$
Number of spikelets per spike	-0.027; $p = 0.849$	-0.031; $p = 0.831$
Plant height	-0.450; $p < 0.000$	-0.481; $p = 0.0004$
1000 kernel weight	-0.235; $p = 0.096$	-0.199; $p = 0.160$
Log 10 Zn concentration	0.224; $p = 0.114$	0.282; $p = 0.044$
Log 10 plant weight	-0.533; $p = <0.001$	-0.457; $p = 0.0007$
Log 10 number of spikelets per spike	-0.026; $p = 0.855$	-0.031; $p = 0.83$
Log10 plant height	-0.467; $p = 0.0006$	-0.481; $p = 0.0004$
Log 10 1000 kernel weight	-0.233; $p = 0.099$	-0.199; $p = 0.160$

In conclusion, important step in determination of relationship between variables is selection of the most appropriate measure of correlation. For instance Pearson's r correlation coefficient is parametric measure of correlation that depicts linear relationship between variables unlike to Spearman's rank correlation coefficient that is non-parametric measure of correlation, calculated on ranks and it depicts a monotonic relationship. The most influential factors affecting the choice of correlation coefficient are data type, linearity of relationship, presence of outliers and violation of parametric assumptions. Without doubt coefficient of determination is more informative compared to coefficient of correlation, and should be reported along with correlation coefficient, sample size and probability value. Lastly, results of correlation analysis should be interpreted in a light of fact that correlation does not imply causation while deeper insight could be achieved by means of regression analysis.

ACKNOWLEDGEMENTS

The paper is a result of the research project „Soil conditioning impact on nutrients and heavy metals in soil-plant continuum" (079-0790462-0450) financed by The Ministry of Science, Education and Sports, Croatia.

REFERENCES

1. Artusi, R., Verderio, P., Marubini, E. (2002): Bravais-Pearson and Spearman correlation coefficients: meaning, test of hypothesis and confidence interval. The International Journal of Biological Markers, 17(2): 148-151.
2. Congelosi, V.E., Taylor, P.E., Rice, P.F. (1976): Basic statistics: A real world approach. West publishing Co, Minnesota, USA.
3. Denis, J.D. (2001): The Origins of Correlation and Regression: Francis Galton or Auguste Bravais and the Error Theorists? History and Philosophy of Psychology Bulletin,13: 36–44.
4. Eđed, A., Horvat, D., Lončarić, Z. (2009): Optimal sample size for statistical analysis of winter wheat quantitative traits. Poljoprivreda/Agriculture, 15(1): 34-38.
5. Eđed, A., Lončarić, Z., Horvat, D., Skala, K. (2010): Visualization of winter wheat quantitative traits with parallel coordinate plots. Poljoprivreda/Agriculture, 16(2): 14-19.
6. Hauke, J., Kossowski, T. (2011): Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Questiones Geographicae, 30(2): 97-93.
doi: http://dx.doi.org/10.2478/v10117-011-0021-1
7. Taylor, R. (1990): Interpretation of the correlation coefficient: A basic review. Journal of Diagnostic Medical Sonography, 1: 35-39.
doi: http://dx.doi.org/10.1177/875647939000600106
8. Udovičić, M., Baždarić, K., Bilić-Zulle, L., Petrovečki, M. (2007): What we need to know when calculating the coefficient of correlation? Biochemia Medica, 17(1): 1-138.
doi: http://dx.doi.org/10.11613/BM.2007.002

PEARSONOV ILI SPEARMANOV KOEFICIJENT KORELACIJE – KOJI KORISTITI?

SAŽETAK

Za opisivanje veze između dvaju svojstava najčešće se upotrebljavaju Pearsonov i Spearmanov koeficijent korelacije. Cilj je ovoga rada usporediti vrijednosti Pearsonovog i Spearmanovog koeficijenta korelacije na istom setu podataka. Podatci su prikupljeni iz pokusa provedenog u polukontroliranim uvjetima, u kojemu je ispitivana sortna specifičnost 51 genotipa ozime pšenice s obzirom na akumulaciju Cd i Zn u zrno na nekontaminiranome i Cd kontaminiranome tlu. Ispitivana je veza između koncentracije Cd u zrnu te koncentracije Zn u zrnu, visine biljke, mase biljke, broja klasića po klasu i mase 1000 zrna na nekontaminiranome i Cd kontaminiranome tlu. Na temelju rezultata, možemo zaključiti da izbor odgovarajućega koeficijenta korelacije najviše ovisi o vrsti podataka, prisustvu netipičnih vrijednosti i ispunjavanju parametrijskih pretpostavki.

Ključne riječi: linearna veza, outlieri, log 10 transformacija, raspodjela učestalosti, ozima pšenica

(Received on 15 May 2015; accepted on 27 November 2015 - *Primljeno 15. svibnja 2015.; prihvaćeno 27. studenoga 2015.*)