

# Original Research Article

## **A comparison of the Pearson, Spearman rank, and Kendall tau correlation coefficients using quantitative variables**

### **ABSTRACT**

In all fields and branches of sciences especially statistics, the correlation coefficient is one of the most often used statistical measures. This study has been carried out for comparing the performances of the Pearson ( $r_p$ ), Spearman's Rank ( $r_s$ ), and Kendall's Tau ( $r_K$ ) correlation coefficients under three sample sizes based on the data of quantitative variables of cotton. Descriptive statistics showed the presence of genetic variability for the cotton studied traits in this study. The quantity, significance, and direction of the correlation calculated by  $r_p$  differed in some cases from the other methods under the three sample sizes, opposite is true for  $r_s$  and  $r_K$ . The highest number of positive correlations among studied traits were by  $r_p$  under  $N = 30$  observations, and by  $r_s$  and  $r_K$  under  $N = 20$  observations. The studied correlation methods performances by Root Mean Square Error (RMSE) revealed that  $r_p$  and  $r_s$  appear to be a good estimator of correlation because they have the lowest values of RMSE. The highest values of RMSE were observed by  $r_p$  and  $r_K$  under  $N=10$  and  $N=20$ , and by  $r_K$  under  $N=30$ . The results of PCA could be useful and appropriate in this study, in which the PCA1 had highly positively correlated with the three studied methods for  $N=10$  observations, and with  $r_s$  and  $r_K$  for  $N=20$  observations.

**Key words:** Correlation, Pearson, Spearman's Rank, Kendall's Tau, RMSE, PCA.

### **INTRODUCTION**

The idea of correlation is regarded as a beginning point for the development of many fields of statistical research in theory (Sobri et al., 2016). Correlation analysis is a common statistical approach for determining the direction and degree of a linear relationship between two variables under investigation in all branches of statistics and other science (Etaga et al., 2021). Sir Francis Galton influenced Karl Pearson substantially in his efforts to organize the use of correlation, and in 1896 Pearson produced the current form of the Pearson's Product Moment correlation coefficient (Sobri et al. 2016). The Pearson correlation coefficient, as well as other forms of correlation coefficients, were developed by Pearson (Coblick, 1998).

The use of various correlation coefficients for the same set of data may lead to significantly different conclusions (Temizhan et al. 2022). Despite the fact that many various correlation coefficients have been presented, scientists and researchers most usually utilize the Pearson Product-Moment correlation, despite its lack of robustness (Wilcox, 2012; Tuğran et al. 2015). The application of these correlation approaches to real-world data is contingent on the method's underlying assumptions. As a result, the Pearson correlation coefficient is the most often employed estimate in studies of linear associations between two variables, and it works under the assumptions of continuity, linearity, and normality (Etaga et al., 2021). In actuality, the coefficient can be calculated without any assumptions as a measure of a linear relationship (Schober et al. 2018). The assumption of normality is generally met and thus when it is not met, using Pearson's or Spearman's correlations can lead to serious errors (Temizhan et al. 2022). Many scholars believed that when data anomalies exist, the Pearson method's

efficiency is reduced (McCallister, 1991). When the assumptions of Pearson correlation are not fully satisfied, Pearson alternatives have been offered, such as Spearman-Rank, Kendall-Tau, Median, Quadrant correlation approaches, and so on (Etaga et al., 2021). Pearson's, Spearman Rank, and Kendal-Tau correlation coefficients appear to be the most potent coefficients when all experimental conditions are analyzed collectively (Temizhan et al. 2022).

The objective of this study is to compare the performances of three correlation coefficients namely the Pearson, Spearman's Rank, and Kendall's Tau coefficients under three sample sizes based on the data of quantitative variables. Data of cotton yield and yield components were used to demonstrate the methodology.

## MATERIALS AND METHODS

### Experimental Data:

In this study, the data from El-Hashash (2017) will be used and analyzed. The data on an individual plant basis of the cotton genotypes recorded for the number of bolls/plant (No. of B/P), boll weight in grams (g; BW), seed index (g; SI), lint percentage (L%), seed cotton yield/plant in grams (g; SCY/P), 2.5% Span length (mm; 2.5% SL), fiber fineness (FF) and fiber strength (gm/tex; FS) traits. The sample size of the simulated data consists of 10, 20, and 30 observations. The data will be analyzed using Pearson, Spearman Rank, and Kendall Tau Correlation Coefficients for these traits. To comparison between these methods, the performances of these methods are measured using the criteria of Root Mean Square Error (RMSE) and principle component analysis (PCA). The PCA was performed using a computer software program Origin Pro 2021 version b 9.5.0.193.

### Pearson Correlation Coefficient:

The Pearson correlation coefficient is one of the most extensively used and arguably best-known association measurements among researchers and scientists (Temizhan et al., 2022). The Pearson correlation coefficient (also known as the Pearson product-moment correlation) is a parametric measure of the linear relationship between two numeric variables and is denoted by ( $r_p$ ). It is defined as the ratio of the covariance between the two variables predictor (X) and response (Y) under investigation to the product of their individual standard deviations, with a range of +1 to -1 inclusive (Carroll, 1961; Chung and Lee, 2001; Sobri et al., 2016).

Let X and Y be random variables where,  $x_j, y_j, j = 1, 2, \dots, k$  be the observed bivariate data points. Then the Pearson correlation coefficient (Pearson, 1920) is calculated mathematically from the equation is described as follows:

$$r_p = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Where  $\bar{X}$  and  $\bar{Y}$  are the averages of the X and Y measurements.

### Spearman Rank Correlation Coefficient:

Spearman rank correlation coefficient is a nonparametric measure of the monotonic association to investigate the linear relations between two variables, denoted by  $r_s$  (Spearman, 1904). The sample of  $r_s$  is calculated in the same manner as  $r_p$ , except that  $r_s$  is calculated after both X and Y have been ranked and transformed to values between 1 and N (de Winter et al., 2016). Which is calculated by converting random variables  $X_i$  and  $Y_i$  into ranked variables  $r_{xi}$  and  $r_{yi}$ , respectively (Etaga et al., 2021). The  $r_s$  can be used to assess the monotonic relations based on the rank of the

observations, non-normally distributed data, the sample size is small, with ordinal data, and there is an outlier problem in the data set (Temizhan et al., 2022). The  $r_s$  assumptions are that as the two variables are measured on an ordinal scale, the scores on one must be monotonically related to the other, and there are no ties observations (Etaga et al., 2021; Temizhan et al., 2022).

Let  $d = r_{xi} - r_{yi}$ , denoted the difference between the ranks of the  $i^{th}$  observations in the two variables X and Y. It is assumed that there is no tie, then each of the variable X and Y takes the rank values  $1, 2, \dots, n$ . The rs can be computed by using following equation:

Where  $d_i^2$ : square of the difference between the ranks of the  $i^{th}$  observations X and Y and n: the number of observations (sample size).

#### **Kendall's tau correlation coefficient:**

Kendall's tau ( $\tau$ ) is a non-parametric measure of the association based on the difference between the probabilities of concordance and discordance between two observed variables, X and Y (Kendall 1955). Kendall's tau ( $\tau$ ) is commonly used as a distribution-free measure of cross-correlation between two variables (Kendall and Gibbons, 1990). Various relations are known to hold between the rank correlation coefficients ( $\tau$  and  $r_s$ ) and the correlation  $r_p$  in samples from a normal bivariate population (Kendall, 1949). Kendall's Tau correlation coefficient is used to examine the relationship between two ordinal variables, and it has the same assumptions as the Spearman rank correlation (Etaga et al. 2021; Temizhan et al., 2022).

If C is the number of concordant pairs (are how many larger ranks are below a certain rank in the column under consideration), D is the number of discordant pairs (are how many smaller ranks are below a certain rank in the column under consideration), and n is the sample size, then there will be  $k=n(n-1)/2$  possible comparisons between any pair of rank ( $X_i, Y_i$ ) and any pair of rank ( $X_j, Y_j$ ). The Kendall's Tau correlation can be calculated using the following calculation based on this information (Kendall, 1938):

$$\tau_{tau} = \frac{C - D}{n(n-1)/2} = \frac{2(C - D)}{n(n-1)}$$

The values of Kendall's tau ranged between -1 and +1 (Sheskin, 2011).

#### **Root Mean Square Error (RMSE):**

The RMSE is defined as the square root of the variance of an estimated value  $\hat{\rho}$  of true value  $\rho$  (Etaga et al. 2021), or it is an estimator with respect to the estimated parameter is defined as the square root of the mean square error (Sobri et al. 2016). The RMSE is a frequently used measure of the differences between values predicted by a model or an estimator and the value actually observed from the thing being modeled or estimated. RMSE is a good measure of precision and these individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power to the accuracy of an estimator (Sobri et al. 2016; Etaga et al. 2021). Calculation of the RMSE as follows:

$$\begin{aligned} \text{Deviation of estimated value from its true value} &= (\hat{\rho} - \rho) \\ \text{Square of deviation} &= (\hat{\rho} - \rho)^2 \\ \text{Variance} &= E((\hat{\rho} - \rho)^2) \end{aligned}$$

$$RMSE(\hat{\rho}) = \sqrt{MSE(\hat{\rho})} = \sqrt{E((\hat{\rho} - \rho)^2)}$$

The better correlation method is thought to be one that has a low RMSE value (Etaga et al. 2021).

## RESULTS AND DISCUSSION

### Descriptive statistics:

Descriptive statistics for studied traits of cotton under different sample size combinations (10, 20 and 30) are presented in Table 2. Under sample sizes of 10, 20, and 30, the values of range, mean, and standard error revealed differences for all analyzed traits, indicating the presence of genetic variability for the studied traits of cotton. In plant breeding and genetics, the study of distribution using skewness and kurtosis provides information about the nature of gene action type (Fisher et al. 1932) and the number of genes controlling the traits (Robson, 1956). The traits including BW, SI, and FS for the three sample size combinations as well as the SCY/P trait for N = 20 showed positive skewness which indicates complementary gene action, while other studied traits had negative skewness under three sample sizes, suggesting duplicate gene effects. SCY/P, L%, and fiber traits (2.5%SP, FS, and FF) for N = 10 as well as L% and FS traits for other sample sizes exhibited gene interactions, due to the positive Kurtosis for these traits. On the other hand, the other traits under different sample size combinations demonstrated an absence of gene interactions, due to the negative Kurtosis for these traits. The above results are similar to findings that were observed in earlier studies such as Nachimuthu et al. (2014), Ponnaiah et al. (2017) and Govintharaj et al. (2018).

The coefficient of variation (CV%) was estimated and categorized as very high ( $CV\% \geq 21$ ), high ( $15 \leq CV\% < 21$ ), moderate ( $10 \leq CV\% < 15$ ), and low ( $CV\% < 10$ ) according to Gomes (2009). The CV% values registered for most evaluated traits across three sample size combinations were low ( $CV < 10\%$ ), indicating high precision and reliability of the field experiments carried out, due to the environmental influence was low. The No. of B/P and FF traits for three sample sizes, SCY/P for N = 20 and N = 30, as well as FS traits for N = 10 showed that the CV% values were moderate and greater than that of the other traits measured. Thus, the environmental influence was high for these traits when compared to the other traits. That would suggest the existence of substantial differences for studied traits to sample size response. A similar trend has been reported in cotton by Yehia and El-Hashash (2021) and El-Hashash and Yehia (2021).

Table 1. Descriptive statistics for yield, yield components, and fiber traits in cotton under different sample size combinations.

Sample Size	Traits Statistics	BW	SCY/P	SI	L%	No. of B/P	2.5%SP	FS	FF
		10	Min	2.62	55.34	8.30	33.89	19.71	27.70
Max	3.24		75.18	9.95	41.28	27.50	35.05	11.85	4.42
Mean	2.91		68.81	9.14	38.43	24.53	32.64	9.90	4.03
Std. error	0.06		1.82	0.17	0.65	0.85	0.78	0.32	0.15
Skewness	0.14		-1.42	0.03	-1.03	-0.85	-1.22	0.26	-1.26
Kurtosis	-0.91		2.87	-1.05	2.06	-0.57	0.53	1.14	0.09
CV%	6.79		8.35	5.98	5.32	11.00	7.52	10.08	11.48
20	Min	2.62	54.82	8.00	33.89	19.71	27.70	8.15	3.10
	Max	3.24	84.10	10.18	42.71	27.55	36.10	12.00	4.60
	Mean	2.91	68.63	9.10	38.57	24.20	32.54	10.11	4.04

	Std. error	0.04	1.78	0.13	0.50	0.65	0.52	0.21	0.10
	Skewness	0.22	0.01	0.07	-0.44	-0.34	-0.44	0.04	-0.97
	Kurtosis	-0.90	-0.37	-0.70	0.57	-1.66	-0.61	0.48	-0.28
	CV%	5.85	11.62	6.42	5.77	12.01	7.18	9.44	11.57
30	Min	2.60	54.82	8.00	33.30	19.71	27.70	8.15	3.10
	Max	3.24	84.10	10.18	42.71	27.55	36.10	12.00	4.60
	Mean	2.89	69.07	9.13	38.30	24.39	32.68	10.15	4.00
	Std. error	0.03	1.30	0.10	0.43	0.48	0.41	0.17	0.08
	Skewness	0.25	-0.02	0.23	-0.56	-0.39	-0.41	0.12	-0.79
	Kurtosis	-0.73	-0.11	-0.76	0.16	-1.39	-0.66	0.35	-0.69
	CV%	5.72	10.34	6.22	6.09	10.86	6.88	9.29	11.23

BW: boll weight; SCY/P: seed cotton yield/plant; SI: seed index; L%: lint percentage; No. of B/P: number of bolls/plant; 2.5%SP: 2.5% Span length; FS: fiber strength; FF: fiber fineness.

### Correlation coefficient methods:

The Pearson ( $r_p$ ), Spearman's Rank ( $r_s$ ), and Kendall's Tau ( $r_k$ ) correlation coefficients were used to study the relationship between studied traits under the three sample size combinations. Under 10 sample size (Table 1), SCY/P and L% positively and significantly correlated with SI, No. of B/P ( $p<0.05$ ), and FF ( $p<0.05$ ) using  $r_p$  only, respectively. While  $r_s$  and  $r_k$  showed a positive and no significant correlation among these traits for a 10 sample size.

As for N=20 observations (Table 2), SCY/P positively and significantly correlated with SI ( $p<0.05$ ) and No. of B/P ( $p<0.01$ ) by three correlation methods, with BW, ( $p<0.05$ ) by  $r_p$  and with FF ( $p<0.05$ ) by  $r_s$  and  $r_k$ . Also, positive and significant correlations were observed between SI and No. of B/P and between L% and FF ( $p<0.05$  or  $p<0.01$ ) by the three correlation methods. Finally, BW was a significantly positive association with L% and FF at 0.05 probability level by  $r_s$  and  $r_k$  methods.

Regarding N= 30 observations (Table 3), the traits of BW, L%, FF (0.05 0.01) using three correlation methods as well as BW, SCY, FF (0.05) using  $r_s$  and  $r_k$  showed a significant correlation among them. SCY/P with SI and No. of B/P (0.050.01) using the three correlation methods and with BW (0.05) using  $r_p$  had positive and significant correlations ( $p<0.05$ ). No. of B/P positively and significantly correlated with SI and 2.5% SP at 0.05 probability level using three methods. Also, positive and significant correlations between 2.5% SP and FS at 0.05 probability level were found  $r_s$  and  $r_k$ , but insignificant using  $r_p$ .

These results indicated that the  $r_s$  and  $r_k$  correlations methods were the same for all relationships between the studied traits under the three sample size combinations. Similar correlation coefficients might be produced by different relationships between variables (Schober et al. 2018). Even for large data sets, the significance of  $r_s$  can lead to the significance or non-significance of  $r_p$ , which is compatible with a logical explanation of the difference between the two coefficients (Hauke and Kossowski, 2011).

Table 2. The values of the Pearson ( $r_p$ ), Spearman's Rank ( $r_s$ ), and Kendall's Tau ( $r_k$ ) correlation coefficients for studied traits under a sample size of N=10.

Methods	Traits	BW	SCY/P	SI	L%	No. of B/P	2.5%SP	FS
$r_p$	SCY/P	0.31						
	SI	0.49	0.59					
	L%	0.08	-0.09	-0.51				
	No. of B/P	-0.39	0.67	0.31	-0.19			
	2.5%SP	-0.26	0.07	-0.12	-0.55	0.27		

	FS	-0.09	-0.21	-0.11	-0.43	-0.35	0.26	
	FF	0.00	-0.02	-0.19	0.66	-0.02	-0.29	-0.69
$r_s$	SCY/P	0.24						
	SI	0.53	0.45					
	L%	0.13	-0.08	-0.44				
	No. of B/P	-0.42	0.52	0.27	-0.36			
	2.5%SP	-0.16	0.13	0.09	-0.72	0.13		
	FS	-0.06	-0.37	-0.20	-0.17	-0.42	0.43	
	FF	0.02	0.23	-0.03	0.41	0.03	-0.29	-0.64
$r_k$	SCY/P	0.13						
	SI	0.31	0.29					
	L%	0.04	-0.02	-0.38				
	No. of B/P	-0.31	0.38	0.20	-0.20			
	2.5%SP	-0.11	0.13	0.04	-0.58	0.13		
	FS	-0.07	-0.23	-0.23	-0.05	-0.32	0.30	
	FF	0.02	0.14	0.00	0.28	0.05	-0.16	-0.50

The studied traits key names can be found in Table 1.

Table 3. The values of the Pearson ( $r_p$ ), Spearman's Rank ( $r_s$ ), and Kendall's Tau ( $r_k$ ) correlation coefficients for studied traits under a sample size of N=20.

Methods	Traits	BW	SCY/P	SI	L%	No. of B/P	2.5%SP	FS
$r_p$	SCY/P	0.41						
	SI	0.31	0.49					
	L%	0.31	-0.04	-0.51				
	No. of B/P	-0.12	0.80	0.39	-0.29			
	2.5%SP	-0.41	0.07	0.07	-0.72	0.38		
	FS	-0.19	-0.09	-0.07	-0.34	-0.17	0.24	
	FF	0.34	0.24	-0.14	0.62	-0.03	-0.49	-0.48
$r_s$	SCY/P	0.30						
	SI	0.38	0.50					
	L%	0.40	0.03	-0.38				
	No. of B/P	-0.11	0.84	0.43	-0.25			
	2.5%SP	-0.41	0.08	0.09	-0.80	0.30		
	FS	-0.17	-0.17	-0.13	-0.18	-0.08	0.31	
	FF	0.41	0.41	0.01	0.55	0.17	-0.43	-0.32
$r_k$	SCY/P	0.21						
	SI	0.28	0.35					
	L%	0.26	0.05	-0.32				
	No. of B/P	-0.07	0.64	0.28	-0.16			
	2.5%SP	-0.31	0.06	0.07	-0.61	0.23		
	FS	-0.14	-0.14	-0.15	-0.10	-0.06	0.22	
	FF	0.32	0.29	0.02	0.40	0.13	-0.28	-0.24

The studied traits key names can be found in Table 1.

Table 4. The values of the Pearson ( $r_p$ ), Spearman's Rank ( $r_s$ ), and Kendall's Tau ( $r_k$ ) correlation coefficients for studied traits under a sample size of N=30.

Methods	Traits	BW	SCY/P	SI	L%	No. of B/P	2.5%SP	FS
$r_p$	SCY/P	0.35						
	SI	0.24	0.47					
	L%	0.33	-0.13	-0.57				
	No. of B/P	-0.20	0.79	0.39	-0.36			
	2.5%SP	-0.47	0.15	0.14	-0.75	0.47		
	FS	-0.16	0.03	0.09	-0.44	-0.08	0.31	
	FF	0.38	0.14	-0.24	0.69	-0.11	-0.57	-0.57
$r_s$	SCY/P	0.33						
	SI	0.28	0.51					
	L%	0.40	-0.11	-0.42				

	No. of B/P	-0.19	0.79	0.42	-0.37			
	2.5%SP	-0.46	0.18	0.14	-0.82	0.41		
	FS	-0.16	-0.01	0.02	-0.23	-0.03	0.35	
	FF	0.41	0.30	-0.08	0.56	0.09	-0.50	-0.41
$r_K$	SCY/P	0.22						
	SI	0.20	0.38					
	L%	0.25	-0.09	-0.34				
	No. of B/P	-0.12	0.60	0.26	-0.25			
	2.5%SP	-0.36	0.13	0.07	-0.60	0.29		
	FS	-0.11	-0.03	-0.05	-0.14	-0.03	0.26	
	FF	0.31	0.22	-0.03	0.38	0.08	-0.33	-0.31

The studied traits key names can be found in Table 1.

The highest number of positive correlations among studied traits by  $r_p$  were recorded under  $N = 30$  observations, followed by  $N = 20$  and  $N = 10$  observations. According to Pernet et al. (2013), as long as samples were drawn from the normal distribution, Pearson's correlation was the best method for measuring the genuine effect sizes and demonstrating greater power. The  $r_S$  and  $r_K$  had given the greatest number of positive associations under  $N = 20$  observations, while they had presented the minimum under  $N = 10$  and  $N = 30$ , respectively. Compared with that the  $r_S$  and  $r_K$ , the  $r_p$  had the lowest number of positive correlations under  $N = 10$  and  $N = 20$  observations, but it was the highest across  $N = 30$  observations. According to Hauke and Kossowski (2011), there may be instances where  $r_p$  is negative yet  $r_S$  is positive. Based on three correlation methods, there were more negative and positive significant correlations for all measured traits under  $N = 30$  observations when compared to  $N = 20$  and  $N = 10$  observations. Although switching to  $r_S$  instead of  $r_p$  can result in essential efficiency benefits, the effect of sample size is far more dramatic, hence we encourage researchers to constantly assess the confidence interval of their observed effects (de Winter et al., 2016). When compared to the  $r_S$  and  $r_K$ , the significant correlations of  $r_p$  were slightly lower for 20 and 30 sample sizes, while it was slightly greater for 10 sample size. These results indicated the extent of concordance between  $r_S$  and  $r_K$  three sample sizes. The distribution shapes had a detrimental impact on the  $r_p$ , particularly for small sample sizes. This effect was even more obvious for the  $r_S$  and  $r_K$  correlation coefficients (Tuğran et al. 2015). Humphreys et al. (2019) suggest that corrections for the issue of underestimation in  $r_p$  should not be adopted if either the data deviate from bivariate normality or the sample size is greater than around 30.

The same direction (positive or negative) and significant correlations among all studied traits were found using  $r_S$  and  $r_K$ , but they differed in their consistency in quantity under three sample sizes. Where the values of correlations by  $r_S$  in all cases were higher than the by  $r_K$  at the three sample sizes. Similar results were obtained by de Winter et al. (2016), who reported  $r_p$  and  $r_S$  are about 50% greater than  $r_K$  for typical bivariate normal distributions. On the other hand, the quantity, significance, and direction of the correlation calculated by  $r_p$  differed in some cases from the other methods under the three sample sizes. The  $r_K$  is perhaps even more robust and efficient than  $r_S$  (Croux & Dehon, 2010). According to Xu et al. (2013), the  $r_S$  has a lower computational load than  $r_K$ , and that the variance of  $r_S$  can be approximated with high numerical accuracy, leading the authors to conclude that the mathematical advantage of  $r_K$  over  $r_S$  is not of great importance.

### RMSE application:

The performances of the three correlation methods are measured by Root Mean Square Error (RMSE). The RMSE was computed based on the relation between SCY/P and other traits under the three sample size combinations. With regard to N=10 and N=20, the RMSE of  $r_p$  and  $r_K$  was higher than the RMSE of  $r_s$  for the relationships under study. While the RMSE of  $r_K$  was higher than the RMSE of  $r_p$  and  $r_s$  with respect to N=30. Sample size has an impact on these correlation coefficients' performances (Temizhan et al. 2022). According to the mean of RMSE ranking from high to low for the relationships under study, the  $r_K$  recorded the first rank, followed by  $r_p$  and  $r_s$  for the three sample size combinations. These findings were consistent with Etaga et al. (2021). Generally,  $r_p$  and  $r_s$  appears to be good estimator of correlation because they have the lowest values of RMSE. Similar results were reported by Sobri et al. (2016). 25 sample size produces average errors that are frequently much greater than the absolute magnitude of the correlation coefficient, which essentially means that the observed correlations are almost meaningless (de Winter et al. 2016). Unless the sample size is very small, the issue of sample bias is unlikely to call for substantial modification of study conclusions (Humphreys et al. 2019).

Table 5. The values and ranking (R) of the root mean square error (RMSE) of the three correlation coefficients for the three sample size combinations.

Sample Size	Methods	BW		SI		L%		No. of B/P		2.5%SP		FS		FF		Mean of R
		RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R	
10/20	$r_p$	0.64	3	0.46	3	1.07	1	0.27	3	0.93	1	1.15	3	0.90	1	2.14
	$r_s$	0.73	2	0.53	2	1.03	2	0.36	2	0.9	3	1.27	1	0.69	3	2.14
	$r_K$	0.83	1	0.68	1	0.99	3	0.51	1	0.91	2	1.19	2	0.79	2	1.71
10/30	$r_p$	0.67	3	0.47	3	1.11	1	0.28	3	0.89	1	1.10	3	0.94	1	2.14
	$r_s$	0.72	2	0.52	2	1.10	2	0.37	2	0.85	3	1.20	1	0.74	3	2.14
	$r_K$	0.83	1	0.67	1	1.06	3	0.52	1	0.87	2	1.13	2	0.82	2	1.71
20/30	$r_p$	0.62	3	0.52	2	1.09	1	0.21	2	0.89	2	1.03	3	0.81	1	2.00
	$r_s$	0.69	2	0.50	3	1.04	2	0.19	3	0.87	3	1.09	1	0.65	3	2.43
	$r_K$	0.79	1	0.64	1	1.02	3	0.38	1	0.91	1	1.09	1	0.75	2	1.43
Mean	$r_p$	0.64	3	0.48	3	1.09	1	0.25	3	0.90	1	1.09	3	0.88	1	2.14
	$r_s$	0.71	2	0.52	2	1.06	2	0.31	2	0.87	3	1.19	1	0.69	3	2.14
	$r_K$	0.82	1	0.66	1	1.02	3	0.47	1	0.90	2	1.14	2	0.79	2	1.71

The studied traits key names can be found in Table 1.

### Principal component analysis (PCA):

When constructing bivariate associations and creating a matrix of correlation coefficients to be used with a multivariate statistical technique like principal component analysis, the correlation coefficient choice is crucial (de Winter et al., 2016). Principal component analysis (PCA) was used to assess the relationship between the three correlation methods under the three sample size combinations based on the relationships among traits under study. Table 6 lists the seven PCAs for the three correlation approaches using the three sample size combinations. The PCA1, PCA2 and PCA3 extracted had eigenvalues higher than one (Eigenvalue >1) with values of 2.98, 1.88, and 1.34, respectively, and they account for 88.67% of the total variability of variables. While, the other PCAs had eigenvalues that were smaller than one (Eigenvalue <1), and explain 11.33 of the total variance of variables. The PCA1, PCA2 and PCA3 explained, in that order, 42.60%, 26.93%, and 19.14% of the total variance of variables, respectively. Therefore, under the major effect of the three sample size combinations, the first two PCAs can be used as the foundation for

evaluating the association among these methods. These results are in accordance with the findings of de Winter et al., (2016) and Sharma et al. (2022). The PCA1 showed a positive correlation with the three methods for N=10 observations, and with  $r_S$  and  $r_K$  for N=20 observations. While PCA2 had highly positively associated with  $r_S$  under the three sample sizes. In addition to, the PCA3 is positively correlated with  $r_K$  for N=10, with  $r_S$  and  $r_K$  for N=20, and with three methods for N=30. As a result, the PCA1 and PCA3 can be referred to as the relationship by three methods under a small and large sample sizes, respectively, while PCA2 can be named the relationship by  $r_S$  under a three sample size. Based on the PCA1 and PCA2,  $r_P$  correlation coefficient decrease, while  $r_S$  and  $r_K$  correlation coefficients assume a sizeable negative value (Sharma et al. 2022). In comparison to other correlation coefficients, the three correlation coefficients provided the most consistent results (Temizhan et al. 2022). Based on results by de Winter et al., (2016), the  $r_S$  correlation appears to be applicable across a broad array of normal and non-normal distributions.

Table 6. Results of seven PCAs for the three correlation methods under the three sample size combinations.

Sample Size	Methods	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7
10	$r_P$	0.77	-0.53	-2.60	-0.37	0.41	-0.01	0.04
	$r_S$	1.48	1.20	-0.45	-0.40	-0.70	0.01	0.11
	$r_K$	1.87	-1.30	0.51	-0.58	-0.12	0.14	-0.20
20	$r_P$	-1.24	-0.64	-0.72	1.56	-0.32	-0.13	-0.07
	$r_S$	0.46	2.36	0.39	0.69	0.35	0.03	-0.06
	$r_K$	1.82	-0.88	1.32	0.68	0.25	0.13	0.14
30	$r_P$	-3.12	-1.02	0.23	-0.28	-0.04	0.29	0.06
	$r_S$	-1.61	1.74	0.39	-0.75	0.08	-0.01	-0.04
	$r_K$	-0.44	-0.93	0.93	-0.55	0.10	-0.46	0.03
Eigenvalue		2.98	1.88	1.34	0.62	0.12	0.04	0.01
Variance %		42.60	26.93	19.14	8.82	1.74	0.62	0.16
Cumulative%		42.60	69.53	88.67	97.48	99.22	99.84	100.00

$r_P$ ,  $r_S$  and  $r_K$ : Pearson, Spearman rank, and Kendall tau correlation coefficients, respectively.

The PCA1 and PCA2 were employed to draw a biplot, as well as mainly distributed and distinguished the three correlation methods with respect to the three sample sizes into four groups according to the relationships among traits under study (Fig.1). The first group (G1) was related to the highest PCA1 and PCA2, and includes the  $r_S$  for 10 and 20 sample sizes. While the second group (G2) comprised the  $r_S$  with respect to N=30 observations (the lowest PCA1 and the highest PCA2). The third group (G3) consisted of  $r_P$  for N=20 as well as  $r_P$  and  $r_K$  for N=30 (the lowest PCA1 and PCA2). The other methods are grouped in the fourth group (G4) with the highest PCA1 and lowest PCA2. The PCA scree plot for the three correlation methods with respect to three sample size combinations on the relationships under this study showed that the PCA1 and PCA2 eigenvalues correspond to the whole percentage of the variance in the dataset (Fig. 1). The results of the scree plot were harmonic with El-Hashash (2016), Yehia and El-Hashash (2021), El-Hashash et al. (2022) and El Sherbiny et al. (2022), who reported that there is a break in the plot that separates the meaningful components from the trivial components. Thus, most researchers would agree that PC1 and PC2 are probably meaningful.

The  $r_S$  and  $r_K$  correlation estimators combine a bounded and smooth influence function with high statistical efficiency while being fairly robust (Croux and Dehon, 2010). It appears unlikely that  $r_K$  could replace  $r_P$  because modern researchers are

accustomed to interpreting  $r_p$ , while  $r_s$  has the potential to be used in place of  $r_p$ , because,  $r_s$  can surpass  $r_p$  in estimating the population  $r_p$  correlation coefficient (de Winter et al., 2016). In line with the findings and conclusion of Ahad et al. (2018) and Etaga et al. (2021), the  $r_p$  method is appropriate to adopt when data is perfect data (non-contaminated), while for contaminated data, the  $r_s$  method should be used, followed by  $r_K$  method. Li et al. (2012) mentioned that  $r_K$  can be used to replace  $r_p$  correlation when data is not normally distributed with a linear relationship.

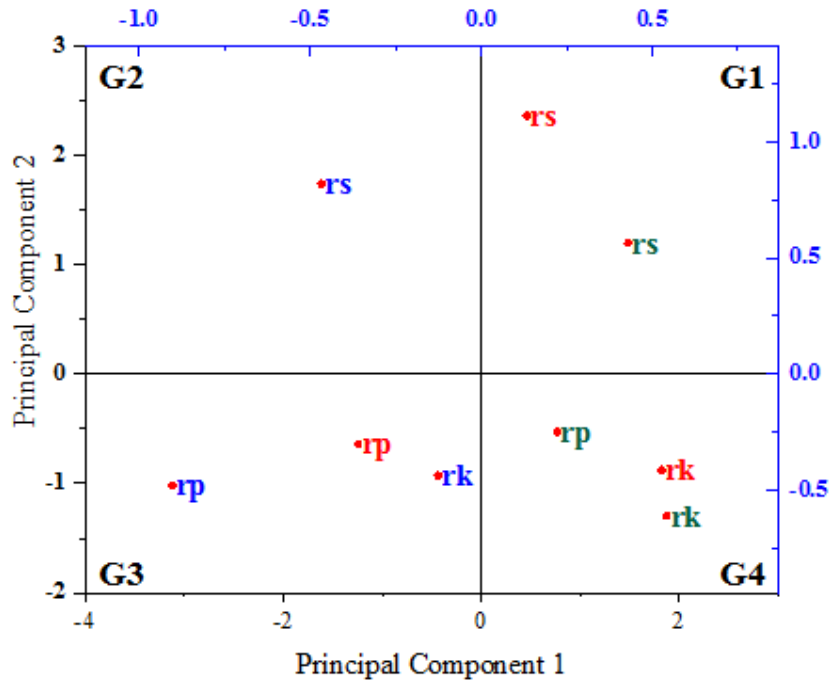


Fig. 1. The biplot diagram based on PCA1 and PCA2 shows the three correlation methods with respect to the three sample sizes (10, 20 and 30).  $r_p$ ,  $r_s$  and  $r_K$ : Pearson, Spearman rank, and Kendall tau correlation coefficients, respectively. 10, 20, and 30 sample sizes are green, red, and blue colors, respectively.

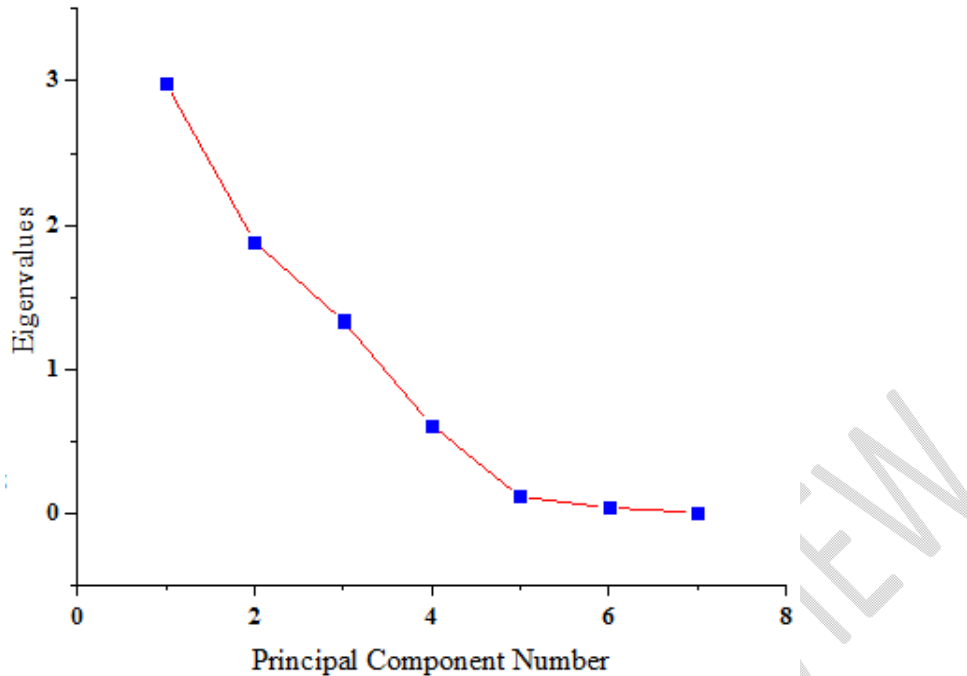


Fig. 2. Scree plot of PCA between respective eigenvalues % and components number.

### CONCLUSION

According to the relationships, RMSE and PCA, all methods seem to work quite well for calculating correlation for quantitative variables with respect to a 30 sample size. Where these results indicated the relationships determined by these methods are most stable and or close to each other when the sample size is large, the opposite is true. Using this information and the associated findings, it can be noted that  $r_s$  and  $r_k$  were equal to each other in direct and significant relationships among quantitative traits under the three sample sizes.

### REFERENCES

- Ahad, N. A., Zakaria, N. A., Abdullah, S., Syed Yahaya, S. S., & Yusof, N. (2018). Robust Correlation Procedure via Sn Estimator. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 10(1-10), 115–118. Retrieved from <https://jtec.utem.edu.my/jtec/article/view/3801>
- Carroll, J.B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, 26, 347–372.
- Chung Y.M. and J.Y. Lee. (2001). A corpus-based approach to comparative evaluation of statistical term association measures. *J. Am. Soc. Inform. Sci. & Tec.* 52: 283–96 (2001).
- Coblick ,W. (1998). *Studies in the History of Statistics Method*, London: Arno Press.
- Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods and Applications*, 19, 497–515. <http://dx.doi.org/10.1007/s10260-010-0142-z>
- de Winter, J. C. F., Gosling, S. D., and Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, 21(3), 273–290. <https://doi.org/10.1037/met0000079>

- El-Hashash, E.F. (2017). Comparison of Variance Components Methods for One Way Random Effects Model in Cotton. *Asian Journal of Advances in Agricultural Research*, 3(1), 1-9. <https://doi.org/10.9734/AJAAR/2017/36955>
- El-Hashash, E.F. Genetic Diversity of Soybean Yield Based on Cluster and Principal Component Analyses. *Journal of Advances in Biology & Biotechnology* 2016, 10, 1-9. <https://doi.org/10.9734/JABB/2016/29127>
- El-Hashash EF, Abou El-Enin MM, Abd El-Mageed TA, Attia MAE-H, El-Saadony MT, El-Tarabily KA, Shaaban A. Bread Wheat Productivity in Response to Humic Acid Supply and Supplementary Irrigation Mode in Three Northwestern Coastal Sites of Egypt. *Agronomy*. 2022; 12(7):1499. <https://doi.org/10.3390/agronomy12071499>
- El-Hashash, E.F. and Yehia, W.M.B. 2021. Estimation of heritability, genes number and multivariate analysis using non- segregation and segregation generations in two cotton crosses. *Asian J. of Biochemistry, Genetics and Molecular Biology*, 9(3), 45-62. <https://doi.org/10.9734/ajbgmb/2021/v9i330221>
- El Sherbiny HA, El-Hashash EF, Abou El-Enin MM, Nofal RS, Abd El-Mageed TA, Bleih EM, El-Saadony MT, El-Tarabily KA, Shaaban A. Exogenously Applied Salicylic Acid Boosts Morpho-Physiological Traits, Yield, and Water Productivity of Lowland Rice under Normal and Deficit Irrigation. *Agronomy*. 2022; 12(8):1860. <https://doi.org/10.3390/agronomy12081860>
- Etaga H.O., Okoro I., Aforika K.F., Ngonadi L.O. (2021), Methods of Estimating Correlation Coefficients in the Presence of Influential Outlier(s). *African Journal of Mathematics and Statistics Studies* 4(3), 157-185. DOI: 10.52589/AJMSS-LLNZXUOZ.
- Fisher RA, FR Immer, O Tedin (1932). The genetical interpretation of statistics of the third degree in the study of quantitative inheritance. *Genetics*, 17, 2, 107–124, <https://doi.org/10.1093/genetics/17.2.107>
- Gomes, F.P. *Curso de estatística experimental*. 15.ed. Piracicaba: Esalq, 2009. 477p.
- Govintharaj P., Manonmani S., S. Robin (2018). Variability and genetic diversity study in an advanced segregating population of rice with bacterial blight resistance genes introgressed. *Ciência e Agrotecnologia*, 42(3):291-296. <https://doi.org/10.1590/141370542018423022317>
- Hauke, J. & Kossowski, T. (2011). Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, 30(2) 87-93. <https://doi.org/10.2478/v10117-011-0021-1>
- Humphreys, R.K., Puth, MT., Neuhäuser, M. et al. Underestimation of Pearson's product moment correlation statistic. *Oecologia* 189, 1–7 (2019). <https://doi.org/10.1007/s00442-018-4233-0>
- Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30, pp. 81-93.
- Kendall, M. G. (1949). Rank and product-moment correlation. *Biometrika*, 36, 177-193. <https://doi.org/10.2307/2332540>
- Kendall, M.G., (1955). *Rank Correlation Methods*. New York: Hafner Publishing Co.
- Kendall, M. G., and Gibbons, J. D., 1990. *Rank Correlation Methods*. Fifth edn. London: Griffin.
- Li G., H. Peng, J. Zhang, L. Zhu (2012). Robust Rank Correlation Based Screening. *The Annals of Statistics*, 40(3): 1846-1877. DOI: 10.1214/12-AOS1024.
- McCallister, C. (1991). Phi. Rho. P.M.. Biserial and Point-Biserial “r”: A review of linkages. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.

- Nachimuthu V.V., S. Robin, D. Sudhakar, M. Raveendran, S. Rajeswari and S. Manonmani (2014). Evaluation of rice genetic diversity and variability in a population panel by principal component analysis. *Indian Journal of Science and Technology*, 7(10):1555-1562, 2014. DOI: 10.17485/ijst/2014/v7i10.14
- Pearson, K. (1920). Notes on the history of correlation. *Biometrika* 13(1),25-45.
- Pernet, C.R., Wilcoxon, R.R., & Rousselet, G.A. (2013). Robust correlation analyses: false positive and power validation using a new open source Matlab toolbox. *Frontiers in Psychology*, 3, 1-18. <https://doi.org/10.3389/fpsyg.2012.00606>
- Ponnaiah, G., S. Tannidi; Manonmani, S.; Robin, S. (2017). Estimates of genetic variability, heritability and genetic advance for blast resistance gene introgressed segregating population in rice. *International Journal of Current Microbiology and Applied Science*, 5(12):672-677, 2016. DOI: 10.20546/ijcmas.2016.512.075
- Robson, D.S. (1956). Application of K4 statistics to genetic variance component analysis. *Biometrics*, 12(4):433-444. <https://doi.org/10.2307/3001682>
- Schober, P.; Boer, C.; Schwarte, L.A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*: 126(5):1763-1768. doi: 10.1213/ANE.0000000000002864
- Sharma, R., Mukherjee, A. & Jassal, H.K. (2022). Reconstruction of latetime cosmology using principal component analysis. *The European Physical Journal Plus*, 137, 219 (2022). <https://doi.org/10.1140/epjp/s13360-022-02397-0>
- Sheskin, D. (2011). *Handbook of Parametric and Nonparametric Statistical Procedure* (5th ed.). Boca Raton, FL: CRC Press.
- Sobri N.B.M., Midi H., Ibrahim N.B., Ismail, N.A., Yaacob W.F.W. and Malik M.A.A. (2016) Differences between pearson's product moment correlation coefficient and an absolute value correlation coefficient in the presence of outliers. *Journal of Mathematics and Computing Science*, 1 (1):1-11.
- Spearman, C. (1904). The proof and measurement of association between two things, 15,72-101. *The American Journal of Psychology*, 100(3/4), special centennial issue (Autumn Winter,1987), 441-471. DOI: 10.2307/1422689.
- Temizhan E., H. Mirtagioglu and M. Mendesc (2022). Which Correlation Coefficient Should Be Used for Investigating Relations between Quantitative Variables? *American Scientific Research Journal for Engineering, Technology, and Sciences*, 85(1): 265-277.
- Tuğran, E. , Kocak, M. , Mirtagioglu, H. , Yiğit, S. and 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, 87-101. doi:10.4236/jdaip.2015.33010.
- Wilcoxon R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*, 3rd Edn Oxford: Academic Press.
- Xu, W., Hou, Y., Hung, Y. S., & Zou, Y. (2013). A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. *Signal Processing*, 93, 261-276. <http://dx.doi.org/10.1016/j.sigpro.2012.08.005>
- Yehia, W.M.B.; El-Hashash E.F. Correlation and multivariate analysis across non-segregation and segregation generations in two cotton crosses. *Egyptian Journal of Agricultural Research* 2021, 99, 354-364. DOI: 10.21608/EJAR.2021.81571.1117.