METHODS OF ESTIMATING CORRELATION COEFFICIENTS IN THE PRESENCE OF INFLUENTIAL OUTLIER(S)

Etaga Harrison O., Okoro Ifeanyichukwu,

Aforka Kenechukwu F. and Ngonadi Lilian O.

Department of Statistics, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

Email: ho.etaga@unizik.edu.ng

Cite this article:

Etaga H.O., Okoro I., Aforka 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.

Manuscript History

Received: 17 Nov 2021

Accepted: 10 Dec 2021

Published: 20 Dec 2021

ABSTRACT: Correlation methods are indispensable in the study of the linear relationship between two variables. However, many researchers often adopt inappropriate correlation methods in the study of linear relationships which usually leads to unreliable results. Recurrently, most researchers ignorantly employ the Pearson method in a dataset that contained outliers, instead of more appropriate correlation methods such as Spearman, Kendall Tau, Median and Quadrant which might be suitable in the calculation of correlation coefficient in the presence of influential outliers. It is noted that the accuracy of estimation of correlation coefficients under outliers has been a long-standing problem for methodological researchers. This is due to low knowledge of correlation methods and their assumptions which have led to inappropriate application of correlation methods in research analysis. Five different methods of estimating correlation coefficients in the presence of influential outlier (contaminated data) were considered: Pearson Correlation Coefficient, Spearman Correlation Coefficient, Kendall Tau Correlation Coefficient, Median Correlation Coefficient and Quadrant Correlation Coefficient.

KEYWORDS: Correction, Influential Outlier, Contaminated Data, Pearson, Spearman, Kendall, Quadrant

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

INTRODUCTION

A major statistical technique used to determine the direction and degree of the linear relationship between two variables under study is correlation analysis. Correlation analysis can be defined as a bivariate analysis that measures the strength (degree) and direction of the linear relationship between two variables, Washington (2010). A correlation study tends to find out whether an increase or decrease in one variable corresponds to an increase or decrease in the other variable. The strength of the linear relationship between any two variables can vary from strong, weak to none.

The concept of correlation was introduced by Galton (1889), who later established the beginning of Modern Statistics. Karl Pearson extended the concepts of correlation and the normal curve. He developed the Pearson correlation coefficient and other types of correlation coefficients (Coblick, 1998). Correlation is utilized in educational and psychological research, either as a primary mode of analysis in which major hypotheses are tested or as part of a secondary analysis, providing background information about the linear relationships between variables of interest prior to or following a complex statistical analysis. The application of correlation has been justified in many situations but it has been misused also. There exist statistical models in the literature that are misinterpreted. Correlation is measured by a coefficient which Washington (2010) defined as a number that represents the level of the linear relationship between the two variables under consideration. The coefficient varies between -1 to +1. The correlation coefficient range is represented as:

```
 |--------------------------|--------------------|-----------------------------|-----------------------|
-1.00                    -0.50                 0                          +0.50                +1.00
Strong – ve relationship                                                  Strong + ve relationship
```

There are many correlation methods available to be used in a correlation study such as Pearson, Spearman, Kendall, Median, Quadrant, etc. The adoption of these correlation methods in real-life data depends on the underlying assumptions peculiar to that method. Hence, the Pearson correlation coefficient is the most widely used estimator in a study of linear association between two variables and it operates under continuity, linearity, and normality assumptions. McCallister (1991), noted that many researchers believed that the effectiveness of the Pearson method is reduced in the presence of data anomaly. However, in practice, these assumptions of Pearson may not be satisfied due to some reasons. This may affect the result of coefficient value and lead to misinterpretation of coefficient value. Therefore, Pearson alternatives have been proposed, such as Spearman-Rank, Kendall-Tau, Median, Quadrant correlation methods, etc which are appropriate to adopt when assumptions of Pearson correlation are not fully satisfied. The distortion of these assumptions may occur due to the existence of the outlie.

Outliers are data values that do not fit the pattern of the data. That means deviated values or extreme values which can bias the estimate of correlation coefficients (Barnett and Lewis, 1995). In other words, outliers are data values lying apart from the rest of the data values. It can occur as an extreme value either on one variable (X or Y) or both variables (X and Y), which may negatively influence the calculation of the correlation coefficient (Wulder, 1996). That is, a single outlier can strongly affect the result of the correlation coefficient. Moreover, non-check of data for outliers may lead to mistakes in data analysis and its interpretation. Anscombe (1960) observed that outliers can exist in any of these three different ways: Inherent data variability, Measurement error, and Execution error. Inherent data variability includes random variation in a normal distribution, distributions with heavy tails, distributions with larger variance, and mixture distributions. Measurement and execution errors include data entry errors, data extraction errors, experimental planning errors, etc. An outlier is illustrated below

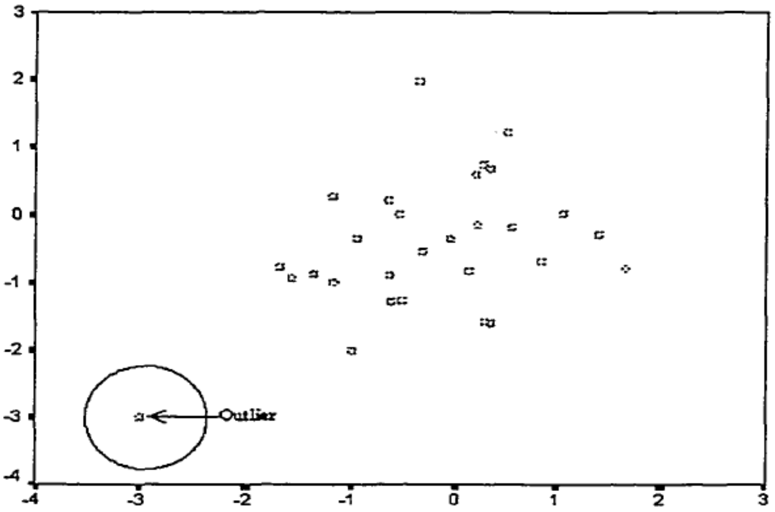

Figure 1.1 Example of Outlier in Sample Data.

In addition, Osbome and Overbay (2004), briefly categorized the deleterious effects of outliers on statistical analyses as follows:

a. Outliers generally serve to increase error variance and reduce the power of statistical tests.

b. Outliers can seriously bias or influence estimates that may be of substantive interest.

c. If distributions are non-randomly, they can decrease normality (as in multivariate analyses) or altering the odds of committing both Type I and Type II errors.

However, it is necessary to study these problems of outliers in correlation analysis as a result of wide spread occurrence of correlation analysis such as correlation and covariance matrices in regression, in multivariate analysis, estimation of the correlation functions etc.

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

LITERATURE REVIEW

This section reviews similar work done by researchers in the past and recent time. It is necessary in order to show the extent of work done by some past researchers and show vividly the gap in knowledge for other researchers to fill.

Mukaka (2012) stated that Pearson correlation coefficient is a parametric statistic and used in normal or approximately normal data only. He further stated that the coefficient value measures the strength of the linear association as the signs denoted the direction of the relationship. He stated that the underlying assumptions are: normality of variables, linearity, homoscedasticity and absence of outliers. He is of the view that, the Spearman correlation coefficient is a nonparametric statistic which is used for data that are not normally distributed or with an unknown distribution.

Abdullah (2015) stated that Pearson correlation performs best under the condition of perfect data (absence of outlier but when data are contaminated with outlier, its performance becomes worst. Also, in simulation studies conducted, he discovered that under perfect correlation coefficient, the performance of product correlation coefficient (s_n) is less compared to median product correlation coefficient (r_m). More so, in terms of average bias and standard error, product correlation coefficient (s_n) performs better when compared to others in most of the condition under study.

Peng et al (2012) observed that, Kendall tau τ rank correlation is a robust correlation measurement between two random variables. He further stated that it can be used to replace Pearson correlation when data is to not normally distributed with a linear relationship.

Leuven (2012) worked on some robust correlations and observed that Spearman and Kendall correlation measures are fairly robust, while maintaining a quite high statistical efficiency.

In a research to compare the performances of a robust correlation coefficient (Median correlation) and the classical correlation coefficient (Person correlation), Abdullah (2015) reported that the median correlation is the best and product correlation (s_n) provides better values if compared to the classical correlation when dataset is contaminated. Other researchers such as Shevlyakov and Smirnov (2011), Sinsomboonthong, (2016), Winter and Gosling, (2016), Anscombe's (1973), Chok (2010), Genest (2003), Fowler (1987), and Tugran et al (2015) have all worked on correlation.

METHODOLOGY

The methods of estimating correlation coefficients under study are presented. Also, criteria for comparison and conditions for simulation are also discussed.

Methods to be compared

Pearson Correlation Coefficient

Pearson correlation is a parametric measures of linear relationship between two numeric variables. It is defined as the ratio of the covariance between the two variables (x, y) under study to the product of their individual standard deviations. It is statistically represented as:

$$r = \frac{Cov(X,Y)}{\sqrt{\sigma_x^2 \sigma_y^2}} \qquad (3.1)$$

$$= \frac{\frac{\sum \quad (X-\underline{X})(Y-\underline{Y})}{n}}{\sqrt{\frac{\sum \quad (X-\underline{X})^2 \sum \quad (Y-\underline{Y})^2}{n \qquad\qquad n}}} \qquad (3.2)$$

$$= \frac{\frac{\sum \quad (X-\underline{X})(Y-\underline{Y})}{n}}{\frac{1}{n}\sqrt{\sum \quad (X-\underline{X})^2 \sum \quad (Y-\underline{Y})^2}} \qquad (3.3)$$

$$= \frac{\sum \quad (X-\underline{X})(Y-\underline{Y})}{\sqrt{\sum \quad (X-\underline{X})^2 \sum \quad (Y-\underline{Y})^2}} \qquad \text{(Gupta,2004)} \qquad (3.4)$$

Where:

X = data values from X variable

Y = data values from Y variable

$\underline{x} = \frac{1}{n_1}\sum$ $\quad x$ = sample mean from X variable

$\underline{y} = \frac{1}{n_2}\sum$ $\quad y$ = sample mean from Y variable

The assumptions of this method include:

a. The two variables (X and Y) are normally distributed
b. There is linear relationship between the two variables (X and Y).
c. The two variables (X and Y) are independent and continuous

Spearman Rank Correlation Coefficient

Spearman rank correlation is a non-parametric measures of the monotonic association between two variables. It is a rank based version of the Pearson correlation. It is calculated by converting random variables X_i and Y_i into ranked variables r_{xi} and r_{yi} respectively.

Assumptions of this method include:

a. The two variables are measured on ordinal scale

b. There are no ties observations

Its sample estimate can be derived as follows:

Let d = $r_{xi} - r_{yi}$, denoted the difference between the ranks of the ith 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, …, n.

This implies that, $\underline{r}_{\text{xi}} = \underline{r}_{\text{yi}} = \frac{n+1}{2}$ and $\sigma^2_{rxi} = \sigma^2_{ryi} = \frac{n^2 -1}{12}$ (3.5)

Since d = $r_{xi} - r_{yi}$ (3.6)

Add $-\underline{r}_{xi} + \underline{r}_{yi}$ to eqn (3.4)

$$= \{(r_{xi} - \underline{r}_{xi}) - (r_{yi} - \underline{r}_{yi})\} \quad (3.7)$$

Square both sides of the eqn in (3.5)

$$d^2 = \{(r_{xi} - \underline{r}_{xi}) - (r_{yi} - \underline{r}_{yi})\}^2 \quad (3.8)$$

Open the bracket on LHS of eqn (3.6)

$$d^2 = (r_{xi} - \underline{r}_{xi})^2 - (r_{yi} - \underline{r}_{yi})^2 - 2(r_{xi} - \underline{r}_{xi})(r_{yi} - \underline{r}_{yi}) \quad (3.9)$$

Summing both side over n values and dividing by n

$$\frac{\sum \quad d^2}{n} = \frac{\sum \quad (r_{xi}-\underline{r}_{xi})^2}{n} + \frac{\sum \quad (r_{yi}-\underline{r}_{yi})^2}{n} - \frac{2\sum \quad (r_{xi}-\underline{r}_{xi})(r_{yi}-\underline{r}_{yi})}{n} \qquad (3.10)$$

$$\frac{\sum \quad d^2}{n} = \sigma^2_{rxi} + \sigma^2_{ryi} - \frac{2\sum \quad (r_{xi}-\underline{r}_{xi})(r_{yi}-\underline{r}_{yi})}{n} \qquad (3.11)$$

But Spearman rank correlation (r_s) is given by

$$r_s = \frac{\sum \quad (r_{xi}-\underline{r}_{xi})(r_{yi}-\underline{r}_{yi})}{n\sigma_{rxi}\sigma_{ryi}} \qquad (3.12)$$

$$r_s\ \sigma_{\mathrm{rxi}}\sigma_{\mathrm{ryi}} = \frac{\sum \quad (r_{xi}-\underline{r}_{xi})(r_{yi}-\underline{r}_{yi})}{n} \qquad (3.13)$$

substituting (3.11) in (3.9)

$$\frac{\sum \quad d^2}{n} = \sigma^2_{rxi} + \sigma^2_{ryi} - 2\ r_s\sigma_{\mathrm{rxi}}\sigma_{\mathrm{ryi}} \qquad (3.14)$$

$$= 2\sigma^2_{rxi} - 2\mathrm{r_s}\sigma^2_{rxi}$$
(3.15)

$$= 2\sigma^2_{rxi}(1 - \mathrm{r_s})$$
(3.16)

$$1 - r_s = \frac{\sum \quad d^2}{2n\sigma^2_{rxi}}$$
(3.17)

$$= \frac{6\sum \quad d^2}{n(n^2-1)}$$
(3.18)

$$r_s = 1 - \frac{6\sum \quad d^2}{n(n^2-1)}$$ (Spearman, 1904). (Gupta,2004) (3.19)

where

d^2 = square of the difference between the ranks of the ith observations X and Y

n = sample size

Kendall Tau Correlation Coefficient

Kendall tau is a non-parametric measures of the association based on concordance and discordance of x-y plane. It has the same assumptions of Spearman method. Its formula is given below.

$$\tau = \frac{(C-D)}{\frac{n(n-1)}{2}} \qquad \text{(Kendall, 1938)} \qquad (3.20)$$

where

n = sample size

C = Concordant pairs (Concordant pairs are how many larger ranks are below a certain rank in the column under consideration)

D = Discordant pairs (Discordant pairs are how many smaller ranks are below a certain rank in the column under consideration).

Median Correlation Coefficient

Median correlation is a non-parametric measure of the association based on median and MAD (median absolute deviation) of X and Y variables. Shafiullah and Khan (2012) derived Median Correlation as follows:

$$r = \frac{\frac{1}{n}\sum \quad (x - \underline{x})(y - \underline{y})}{\sigma_x \sigma_y} \qquad (3.21)$$

$$= \frac{1}{n}\sum \quad \left(\frac{x - \underline{x}}{\sigma_x}\right)\left(\frac{y - \underline{y}}{\sigma_y}\right) \qquad (3.22)$$

$$= \text{mean}(z_x \text{ x } z_y) \qquad (3.23)$$

where $z_x = \frac{x - \underline{x}}{s_x}$ and $z_y = \frac{y - \underline{y}}{s_y}$

Replacing mean with median in eqn (3.24)

$$r_{MED} = \text{median}(Q_x \text{x } Q_y) \qquad (3.24)$$

$$r_{MED} = \text{median}\left(\frac{(x - median(x))}{MAD(x)} \text{ x } \frac{(y - median(y))}{MAD(y)}\right) \qquad \text{(Shafiullah and Khan, 2012)} \quad (3.25)$$

Where

$$Q_x = \frac{x - median(x)}{MAD(x)} \qquad (3.26)$$

$Q_y = \frac{y - median(y)}{MAD(y)}$ are robust standardized variables of X and Y.

$MAD(x)$ = med (| x – med(x) |) stands for median absolute deviation of X

med(x) = med (x_1, x_2, . . .x_n) stands for sample median of X variable.

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

Quadrant Correlation Coefficient

Quadrant correlation is a non-parametric measures of correlation. It makes use of sign function rather than rank of observations. It is statistically represented as:

$$r_Q = \frac{\sum_{i=1}^{n} sign\,(x_i - med(x))sign(y_i - med(y))}{n} \qquad \text{Blomqvist (1950).} \qquad (3.27)$$

Where

n = sample size

Med(x) = median value of X variable

(x – med(x)) = deviation of observations from its median

Sign(x – med(x)) = 1 for positive deviations, -1 for negative deviations and 0 for zero deviations.

Method of Comparison

Simulation studies will be used to compare the properties of selected correlation methods. The following parameters will be used in the simulation study.

Sample Size

Sample size of the simulated data will be varied to include 10, 30 and 50 to cover small and large samples.

Level of Outlier

The levels of outliers in the data will be varied as follows: 1% , 5% and 10%. Outlier will be varied in order to find out the most appropriate correlation method values close to -1 or +1, which is perfect correlation coefficient.

Measures of Performance

The performances of the selected correlation methods are measured under the following criteria: Absolute Bias and Root Mean Square Error (RMSE).

Absolute Bias

Absolute bias of an estimator is defined as the absolute value of difference between the expectation of the estimated values and the true value.

Statistically,

$$\text{ABS}(\hat{\theta}) = |\ \text{E}(\hat{\theta}) - \theta\ | \qquad (3.28)$$

Where

θ = true value

$\hat{\theta}$ = estimated value

E($\hat{\theta}$) = expected value of $\hat{\theta}$

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

The correlation method with smaller value of Bias is considered to be the best correlation methods.

The Root Mean Square Error (RMSE)

The RMSE is defined as the square root of the variance of an estimated value $\hat{\theta}$ of true value θ. The root mean square error measures the accuracy of an estimator. The RMSE is given as

$$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)} \tag{3.29}$$

Where

$(\hat{\theta} - \theta)$ = deviation of estimated value from its true value

$(\hat{\theta} - \theta)^2$ = square of deviation of estimated value from its true value

$E((\hat{\theta} - \theta)^2)$ = variance of estimated value from its true value

Correlation method with a small value of root mean square error is considered to be best correlation method.

RESULTS

R version 3.5.1 was used to simulate contaminated and non-contaminated bivariate distributions. Simulated data were analyzed and presented in tables. Also, real life data were collected from Federal College of Agriculture, Ishiagu,Eboyi State and from textbook titled Fundamentals of Statistics by S.C. Gupta. In order to generate distribution of correlation coefficients for illustration of real life application of correlation methods, Absolute Bias and Root Mean Square Error were computed.

Real Life Data (One)

This section illustrates the application of five different methods of estimating correlation coefficients as discussed in chapter three. The data for this real life application were obtained from Federal College of Agriculture, Ishiagu. Seven students were randomly selected out of twenty-one students in department of fishery, ND 1, the scores of these students in two courses: BAM 101 and COM 111 were given below.

Table 1. Scores of seven students in two courses.

Student	1	2	3	4	5	6	7
BAM 101	65	37	74	40	50	75	85
COM 111	70	40	81	46	60	76	73

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

Application of Correlation Methods on Students Scores

a. Pearson Correlation Coefficient

$$r_p = \frac{\Sigma \quad (X-\underline{X}\ )(Y-\underline{Y})}{\sqrt{\Sigma \quad (X-\underline{X}\ )^2 \Sigma \quad (Y-\underline{Y}\ )^2}}$$

Table 2. Computation for Pearson Correlation Coefficient.

	x	y	x - $\underline{x}$	y - $\underline{y}$	(x - $\underline{x}$)(y - $\underline{y}$)	$(x - \underline{x})^2$	$(y - \underline{y})^2$
1	65	70	4.1429	6.2857	26.0408	17.1633	39.5102
2	37	40	-23.8571	-23.7143	565.7551	569.1633	562.3673
3	74	81	13.1429	17.2857	227.1837	172.7347	298.7959
4	40	46	-20.8571	-17.7143	369.4694	435.0204	313.7959
5	50	60	-10.8571	-3.7143	40.3265	117.8776	13.7959
6	75	76	14.1429	12.2857	173.7551	200.0204	150.9388
7	85	73	24.1429	9.2857	224.1837	582.8776	86.2245
Total	426	446			1626.714	2094.857	1465.429

$$r_p = \frac{1626.714}{\sqrt{2094.857 \times 1465.429}} = 0.9285$$

b. Spearman Rank Correlation Coefficient

$$r_s \ = \ 1 - \frac{6\Sigma \quad d^2}{n(n^2-1)}$$

Table 3. Computation for Spearman Rank Correlation Coefficient.

	x	Y	rank (x)	rank (y)	d = x_r - y_r	d^2
1	65	70	4	4	0	0
2	37	40	1	1	0	0
3	74	81	5	7	-2	4
4	40	46	2	2	0	0
5	50	60	3	3	0	0
6	75	76	6	6	0	0
7	85	73	7	5	2	4
Total						8

$$r_s \ = \ 1 - \frac{6 \times 8}{7(49-1)} = 0.8571$$

c. Kendall Tau Correlation Coefficient

$$r_k = \frac{(\,C-D\,)}{\frac{n(n-1)}{2}}$$

Table 4. Computation for Kendall Tau Correlation Coefficient.

	x	Y	rank (x)	rank (y)	C	D
1	37	40	1	1	6	0
2	40	46	2	2	5	0
3	50	60	3	3	4	0
4	65	70	4	4	3	0
5	74	81	5	7	0	2
6	75	76	6	6	0	1
7	85	73	7	5	0	0
Total					21	3

$$r_k = \frac{18-3}{\frac{7(7-1)}{2}} = 0.7143$$

d. Quadrant Correlation Coefficient

$$r_Q = \frac{\sum_{i=1}^{n} \quad sign\,(x_i - med(x))sign(y_i - med(y))}{n}$$

Table 5. Computation for Quadrant Correlation Coefficient.

	x	y	x – 65	sign(x – 65)	y – 70	sign(y – 70)	sign(x – 65) sign(y – 70)
1	65	70	0	0	0	0	0
2	37	40	-28	-1	-30	-1	1
3	74	81	9	1	11	1	1
4	40	46	-25	-1	-24	-1	1
5	50	60	-15	-1	-10	-1	1
6	75	76	10	1	6	1	1
7	85	73	20	1	3	1	1
Total							6

$$r_Q = \frac{6}{7} = 0.8571$$

e. Median Correlation Coefficient

$$r_{MED} = \text{median}\left(\frac{(x - median(x))}{MAD(x)}\, x\, \frac{(y - median(y))}{MAD(y)}\right)$$

Median(x) = 65

Median(y) = 70

$MAD(x)$ = med (| x – med(x) |) = 15

$MAD(y)$ = med (| y – med(y) |) = 10

Table 6. Computation for Median Correlation Coefficient.

	x	Y	(x – 65)	(y – 70)	$\frac{(x-65)}{15}$	$\frac{(y-70)}{10}$	$\frac{(x-65)}{15} x \frac{(y-70)}{10}$
1	65	70	0	0	0.00	0.00	0.00
2	37	40	–28	–30	–1.87	–3.00	5.60
3	74	81	9	11	0.600	1.10	0.66
4	40	46	–25	–24	–1.67	–2.40	4.00
5	50	60	–15	–10	–1.00	–1.00	1.00
6	75	76	10	6	0.67	0.60	0.40
7	85	73	20	3	1.33	0.30	0.40

r_{MED} = 0.6600

Table 7: The result of real life data

METHODS	COEFFICIENT VALUES
r_p	0.9285
r_s	0.8571
r_k	0.7143
r_m	0.6600
r_q	0.8571

Table 7 present the results of Pearson (r_p), Spearman (r_s), Kendall tau (r_k), Median (r_m) and Quadrant (r_q) correlation coefficients obtained from real life data.

Real Life Data (Two)

In this section, data were secondary data obtained from textbook titled Fundamentals of Statistics by Gupta (2011), Chapter Eight, Exercise 8.4, No 16. It provided data on ten observations in two variables: X and Y as given below

TABLE 8. Source Gupta (2004) exercise 8.4, no 16.

variable	1	2	3	4	5	6	7	8	9	10
X	78	36	98	25	75	82	92	62	65	39
Y	84	51	91	69	68	62	86	58	35	49

Application of Correlation Methods on the Secondary Data

a. Pearson Correlation Coefficient (r_p)

$$r_p = \frac{\Sigma \quad (X - \underline{X}\ )(Y - \underline{Y})}{\sqrt{\Sigma \quad (X - \underline{X}\ )^2 \Sigma \quad (Y - \underline{Y}\ )^2}}$$

Table 9. Computation for Pearson Correlation Coefficient for Real life Data.

	x	Y	x - $\underline{x}$	y - $\underline{y}$	(x - $\underline{x}$)(y - $\underline{y}$)	$(x - \underline{x})^2$	$(y - \underline{y})^2$
1	78	84	12.8	18.70	239.36	163.84	349.69
2	36	51	-29.2	-14.30	417.56	852.64	204.49
3	98	91	32.8	25.7	842.96	1075.84	660.49
4	25	69	-40.2	3.70	-148.74	1616.04	13.69
5	75	68	9.8	2.70	26.46	96.04	7.29
6	82	62	16.8	-.3.30	-55.44	282.24	10.89
7	92	86	26.8	20.70	554.76	718.24	428.49
8	62	58	-3.2	-7.30	23.36	10.24	53.29
9	65	35	-0.2	-30.30	6.06	0.04	918.09
10	39	49	-26.2	-16.30	427.06	686.44	265.69
Total					2333.4	5501.6	2912.1

$$r_p = \frac{2333.4}{\sqrt{5501.6\ x\ 2912.1}} = 0.5829$$

b. **Spearman Rank Correlation Coefficient (r_s)**

$$r_s = 1 - \frac{6\sum \quad d^2}{n(n^2-1)}$$

Table 10. Computation for Spearman Rank Correlation Coefficient Real life data.

	x	Y	rank (x)	rank (y)	d = x_r - y_r	d^2
1	78	84	7	8	-1	1
2	36	51	2	3	-1	1
3	98	91	10	10	0	0
4	25	69	1	7	-6	36
5	75	68	6	6	0	0
6	82	62	8	5	3	9
7	92	86	9	9	0	0
8	62	58	4	4	0	0
9	65	35	5	1	4	16
10	39	49	3	2	1	1
Total						64

$$r_s = 1 - \frac{6 \times 64}{10(10-1)} = 0.6121$$

c. **Kendall Tau Correlation Coefficient (r_k)**

$$r_k = \frac{(C-D)}{\frac{n(n-1)}{2}}$$

Table 11. Computation for Kendall Tau Correlation Coefficient Real life data.

	x	Y	rank (x)	rank (y)	C	D
1	25	69	1	7	3	6
2	36	51	2	3	6	2
3	39	49	3	2	6	1
4	62	58	4	4	5	1
5	65	35	5	1	5	0
6	75	68	6	6	3	1
7	78	84	7	8	2	1
8	82	62	8	5	2	0
9	92	86	9	9	1	0
10	98	91	10	10	0	0
Total					33	12

$$r_k = \frac{33-12}{\frac{10(10-1)}{2}} = 0.4667$$

d. Quadrant Correlation Coefficient

$$r_Q = \frac{\sum_{i=1}^{n} sign\,(x_i - med(x))sign(y_i - med(y))}{n}$$

Table 12. Computation for Quadrant Correlation Coefficient Real life data.

	X	y	x – 70	sign(x – 70)	y – 65	sign(y – 65)	sign(x – 70) sign(y – 65)
1	78	84	8	1	19	1	1
2	36	51	-34	-1	-14	-1	1
3	98	91	28	1	26	1	1
4	25	69	-45	-1	4	1	-1
5	75	68	5	1	3	1	1
6	82	62	12	1	-3	-1	-1
7	92	86	22	1	21	1	1
8	62	58	-8	-1	-7	-1	1
9	65	35	-5	-1	-30	-1	1
10	39	49	-31	-1	-16	-1	1
total							6

$$r_Q = \frac{6}{10} = 0.6000$$

e. Median Correlation Coefficient

$$r_{MED} = \text{median}(\frac{(x - median(x))}{MAD(x)}\, x\, \frac{(\,y - median(y))}{MAD(y)})$$

Median(x) = 70

Median(y) = 65

$MAD(x)$ = med (| x – med(x) |) = 17

$MAD(y)$ = med (| y – med(y) |) = 15

Table 13. Computation for Median Correlation Coefficient Real life data.

	X	Y	(x – 70)	(y – 65)	$\frac{(x-70)}{17}$	$\frac{(y-65)}{15}$	$\frac{(x-70)}{17}x\frac{(y-65)}{15}$
1	78	84	8	19	0.4706	1.2667	0.5961
2	36	51	-34	-14	-2.0000	-0.9333	1.8667
3	98	91	28	26	1.6471	1.7333	2.8549
4	25	69	-45	4	-2.6471	0.2667	-0.7059
5	75	68	5	3	0.2941	0.2000	0.0588
6	82	62	12	-3	0.7059	-0.2000	-0.1412
7	92	86	22	21	1.2941	1.4000	1.8118
8	62	58	-8	-7	-0.4706	-0.4667	0.2196
9	65	35	-5	-30	-0.2941	-2.0000	0.5882
10	39	49	-31	-16	-1.8235	-1.0667	1.9451

r_{MED} = 0.5923

Table 14: The result of real life data

METHODS	COEFFICIENT VALUES
Pearson (r_p)	0.5829
Spearman ($r_{s)}$	0.6121
Kendall (r_k)	0.4667
Median (r_m)	0.5923
Quadrant (r_q)	0.6000

Application of Absolute Bias on Distribution of Correlation Coefficients

Bias is given as:

B($\hat{\theta}$) = E($\hat{\theta}$) – θ

Where θ = the true population parameter

$\hat{\theta}$ = the estimated value.

a. Bias on Pearson correlation coefficients for fixed rho ($\theta = 1.0$)

B($\hat{\theta}$) = E($\hat{\theta}$) – θ

E($\hat{\theta}$) = $\frac{0.9285+0.5829}{2}$ = 0.755

θ = 1.0

B($\hat{\theta}$) = 0.7559 – 1.0

= -0.2441

Then abs(B($\hat{\theta}$)) = 0.2441

b. Bias on Spearman correlation coefficients for fixed rho ($\theta = 1.0$)

B($\hat{\theta}$) = E($\hat{\theta}$) – θ

E($\hat{\theta}$) = $\frac{0.8571+0.6121}{2}$ = 0.734

θ = 1.0

B($\hat{\theta}$) = 0.7346 – 1.0

= -0.2654

Then abs(B($\hat{\theta}$)) = 0.2654

c. Bias on Kendall correlation coefficients for and fixed rho ($\theta = 1.0$)

B($\hat{\theta}$) = E($\hat{\theta}$) – θ

E($\hat{\theta}$) = $\frac{0.7143+0.4667}{2}$ = 0.590

θ = 1.0

B($\hat{\theta}$) = 0.5905 – 1.0

= -0.4095

But abs(B($\hat{\theta}$)) = 0.4095

d. Bias on Quadrant correlation coefficients for fixed rho ($\theta = 1.0$)

B($\hat{\theta}$) = E($\hat{\theta}$) – θ

E($\hat{\theta}$) = $\frac{0.8571+0.6000}{2}$ = 0.7286

θ = 1.0

B($\hat{\theta}$) = 0.7286 – 1.0

= -0.2714

Then abs(B($\hat{\theta}$)) = 0.2714

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

e. Bias on Median correlation for fixed rho ($\theta = 1.0$)

$B(\hat{\theta}) = E(\hat{\theta}) - \theta$

$E(\hat{\theta}) = \frac{0.6600+0.5923}{2} = 0.6262$

$\theta = 1.0$

$B(\hat{\theta}) = 0.6262 - 1.0$

$= -0.3739$

Then abs($B(\hat{\theta})$) = 0.3739

Table 15: The result of computed Absolute Bias on correlation methods and its ranks

Method	Bias	Ranking of Bias (low to high)
Pearson	0.2443	1
Spearman	0.2654	2
Kendall	0.4095	5
Quadrant	0.2714	3
Median	0.3739	4

Table 15 presented the Bias of Pearson (r_p), Spearman (r_s), Kendall tau (r_k), Quadrant (r_Q) and Median (r_m) correlation methods obtained from real life data (one and two).

Application of RMSE on Distribution of Correlation Coefficients

The RMSE is defined as:

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

Where θ is the true population parameter and $\hat{\theta}$ is the estimated value.

i. RMSE on Pearson correlation coefficients for fixed rho ($\theta = 1.0$)

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

Table 16: The result of computed RMSE on Pearson correlation coefficients

	$\hat{\theta}$	θ	$(\widehat{\theta - \theta})$	$(\hat{\theta} - \theta)^2$
1	0.9285	1.0	-0.0715	0.0051
2	0.5829	1.0	-0.4171	0.1740
Total				0.1791

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

$E((\hat{\theta} - \theta)^2) = \frac{0.1791}{2}$

$= 0.0896$

$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}$

$= \sqrt{0.0896}$

$= 0.2913$

ii. RMSE on Spearman correlation coefficients for fixed rho ($\theta = 1.0$)

$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}$

Table 17: The result of computed RMSE on Spearman correlation coefficients

	$\hat{\theta}$	θ	$\widehat{(\theta - \theta)}$	$(\hat{\theta} - \theta)^2$
1	0.8571	1.0	-0.1429	0.0204
2	0.6121	1.0	-0.3879	0.1505
Total				0.1709

$E((\hat{\theta} - \theta)^2 = \frac{0.1709}{2}$

$= 0.0855$

$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}$

$= \sqrt{0.0855}$

$= 0.2924$

iii. RMSE on Kendall correlation coefficients for fixed rho ($\theta = 1.0$)

$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}$

Table 18: The result of computed RMSE on Kendall correlation coefficients

	$\hat{\theta}$	θ	$\widehat{(\theta - \theta)}$	$(\hat{\theta} - \theta)^2$
1	0.7143	1.0	-0.2857	0.0816
2	0.4667	1.0	-0.5333	0.2841
Total				0.3657

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

$$E((\hat{\theta} - \theta)^2 = \frac{0.3657}{2}$$

$$= 0.1829$$

$$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}$$

$$= \sqrt{0.1829}$$

$$= 0.4277$$

iv. RMSE on Median correlation coefficients for fixed rho ($\theta = 1.0$)

$$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}$$

Table 19: The result of computed RMSE on Median correlation coefficients

	$\hat{\theta}$	θ	$\widehat{(\theta - \theta)}$	$(\hat{\theta} - \theta)^2$
1	0.6600	1.0	-0.3400	0.1156
2	0.5923	1.0	-0.4077	0.1662
Total				0.2818

$$E((\hat{\theta} - \theta)^2 = \frac{0.2818}{2}$$

$$= 0.1409$$

$$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}$$

$$= \sqrt{0.1409}$$

$$= 0.3754$$

v. RMSE on Quadrant correlation coefficients for fixed rho ($\theta = 1.0$)

$$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}$$

Table 20: The result of computed RMSE on Quadrant correlation coefficients

	$\hat{\theta}$	θ	$\widehat{(\theta - \theta)}$	$(\hat{\theta} - \theta)^2$
1	0.8571	1.0	-0.1429	0.0204
2	0.6000	1.0	-0.4000	0.1600
Total				0.1804

$$E((\hat{\theta} - \theta)^2 = \frac{0.1804}{2}$$

$$= 0.0902$$

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

$$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}$$

$$= \sqrt{0.0902}$$

$$= 0.3003$$

Table 21: The result of computed RMSE on correlation methods and its ranks

Method	RMSE	Ranking of RMSE (low to high)
Pearson	0.2913	1
Spearman	0.2924	2
Kendall	0.4277	5
Quadrant	0.3003	3
Median	0.3754	4

Simulation of Non-Contaminated Data.

Non-contaminated data were simulated from bivariate normal distribution with means (0.05 and 0.025), standard deviations (0.05 and 0.3), rhos were set at (0.3 and 0.9) and sample sizes (n = 10, 30 and 50) were used. In addition, simulation was replicated three times for each sample sizes in order to generate distribution of correlation coefficients for computation of Absolute Bias and Root Mean Square Error. The result of simulated data in terms of Absolute Bias and Root Mean Square Error were presented in the tables below.

Table 22: The result of simulated non-contaminated data when rho = 0.3, n =10

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.07160	0.19300	1	1
Spearman	0.20333	0.23340	2	2
Kendall	0.21852	0.33776	3	5
Median	0.21780	0.23557	4	3
Quadrant	0.29333	0.28000	5	4

Table 23: The result of simulated non-contaminated data when rho = 0.3, n = 30

Method	Bias	RMSE	Ranking of Bias (high to low)	Ranking of RMSE (high to of low)
Pearson	0.03226	0.15238	1	2
Spearman	0.06752	0.13048	2	1
Kendall	0.12759	0.15922	3	3
Median	0.21374	0.27340	5	5
Quadrant	0.188889	0.25166	4	4

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

Table 24: The result of simulated non-contaminated data when rho = 0.3, n = 50

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.05669	0.07736	1	1
Spearman	0.05917	0.08471	2	2
Kendall	0.08857	0.09237	3	3
Median	0.11342	0.13016	4	4
Quadrant	0.12667	0.14742	5	5

Table 22 to Table 24 gave information about the level of performances of Pearson, Spearman, Kendall tau, Median and Quadrant correlation methods in terms of Absolute Bias and RMSE for rho = 0.3 and sample sizes (n) are 10, 30 and 50.

Table 25: The Result of Absolute Bias and RMSE for rho = 0.9, n =10

Method	Bias	RMSE	Ranking of Bias (high to low)	Ranking of RMSE (high to of low)
Pearson	0.00807	0.01344	1	1
Spearman	0.02626	0.02997	2	2
Kendall	0.10740	0.11529	3	3
Median	0.15963	0.34447	4	4
Quadrant	0.40000	0.42426	5	5

Table 26: The Result of Absolute Bias and RMSE for rho = 0.9, n = 30

Method	Bias	RMSE	Ranking of Bias (high to low)	Ranking of RMSE (high to of low)
Pearson	0.00490	0.04099	1	1
Spearman	0.04638	0.07141	2	2
Kendall	0.20958	0.21751	4	4
Median	0.14806	0.15772	3	3
Quadrant	0.25556	0.28480	5	5

Table 27: The Result of Absolute Bias and RMSE for rho = 0.9, n =50

Method	Bias	RMSE	Ranking of Bias (high to low)	Ranking of RMSE (high to of low)
Pearson	0.01837	0.02123	2	2
Spearman	0.00600	0.01114	1	1
Kendall	0.15959	0.16031	4	4
Median	0.01841	0.05004	3	3
Quadrant	0.16667	0.18294	5	5

Table 25 to Table 27 present results of simulated bivariate normal data when rho = 0.9 and sample sizes (n) are 10, 30 and 50. The performances of Pearson, Spearman, Kendall, Median and Quadrant correlation estimators in regard of Absolute Bias and Root Mean Square Error (RMSE) were also presented.

Simulation of Contaminated Data

In this section, outliers were introduced in the direction of Y distributions from non-contaminated data in section 4.3. Three different levels of outlier: 1%, 5% and 10% were employed. More so, data were contaminated by mixture of normal distribution given below.

$$Y \sim (1 - e)N(\mu\ , \sigma) + (e)N(\mu_1, \sigma_1\ )$$

Where e is outlier level, N($\mu\ , \sigma$) is normal distribution with mean and standard deviation $\mu\ , \sigma$ respectively and N(μ_1, σ_1) is contaminated normal distribution with mean (μ_1) and standard deviation (σ_1). Also so, generated data were replicated three times for each sample sizes in order to generate distribution of correlation coefficients for computation of Absolute Bias and Root Mean Square Error. The result of simulated data in terms of Absolute Bias and Root Mean Square Error were presented in the tables below.

Table 28: The Result of Absolute Bias and RMSE for rho = 0.3, level = 1%, n = 10

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.18511	0.29740	2	3
Spearman	0.20505	0.27880	3	2
Kendall	0.23333	0.26751	4.5	1
Median	0.04208	0.39158	1	5
Quadrant	0.23333	0.30000	4.5	4

Table 29: The Result of Absolute Bias and RMSE for rho = 0.3, level = 1%, n =30

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.21229	0.21870	3	3
Spearman	0.24253	0.24988	4	4
Kendall	0.26245	0.26569	5	5
Median	0.19789	0.20446	2	2
Quadrant	0.14444	0.15753	1	1

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

Table 30: The Result of Absolute Bias and RMSE for rho = 0.3, level = 1%, n = 50

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.27023	0.29630	4	5
Spearman	0.24774	0.25783	2	2
Kendall	0.25510	0.26023	3	3
Median	0.28113	0.28450	5	4
Quadrant	0.23333	0.25377	1	1

Table 28 to Table 30 gave information of contaminated bivariate normal data when rho = 0.3, outlier level = 1% and sample sizes of 10, 30 and 50 respectively. The performances of Pearson, Spearman, Kendall and Quadrant correlation methods under Absolute Bias and (RMSE) were also displayed.

Table 31: The Result of Absolute Bias and RMSE for rho = 0.3, level = 5%, n = 10

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.64951	0.73064	5	5
Spearman	0.60505	0.69022	4	4
Kendall	0.55926	0.62311	3	3
Median	0.41760	0.53591	1	1
Quadrant	0.50000	0.59721	2	2

Table 32: The Result of Absolute Bias and RMSE for rho = 0.3, level = 5%, n = 30

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.33969	0.42425	5	5
Spearman	0.21376	0.25944	1	1
Kendall	0.22817	0.26307	2	2
Median	0.22969	0.37043	3	4
Quadrant	0.27778	0.32375	4	3

Table 33: The Result of Absolute Bias and RMSE for rho = 0.3, level = 5%, n = 50

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.30648	0.31142	5	5
Spearman	0.14421	0.19169	1	1
Kendall	0.19742	0.21376	2	2
Median	0.24056	0.24851	4	4
Quadrant	0.20667	0.22000	3	3

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ

Table 31 to Table 33 contained results of contaminated normal data for fixed rho = 0.3, outlier level 5% and sample sizes (n) were 10, 30 and 50 respectively. The performances of Pearson, Spearman, Kendall, Median and Quadrant correlation methods under Absolute Bias and RMSE were also presented.

Table 34: The Result of Absolute Bias and RMSE for rho = 0.3, level = 10%, n = 10

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.29622	0.46099	4	4
Spearman	0.23333	0.36496	2	1
Kendall	0.24815	0.37024	3	2
Median	0.36855	0.60098	5	5
Quadrant	0.233333	0.55076	1	3

Table 35: The Result of Absolute Bias and RMSE for rho = 0.3, level = 10%, n = 30

Method	Bias	RMSE	Ranking of Bias (high to low)	Ranking of RMSE (high to of low)
Pearson	0.29092	0.33429	5	5
Spearman	0.02317	0.08270	1	1
Kendall	0.07660	0.12635	4	3
Median	0.04357	0.13531	3	4
Quadrant	0.03333	0.11386	2	2

Table 36: The Result of Absolute Bias and RMSE for rho = 0.3, level = 10%, n = 50

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.43281	0.44655	5	5
Spearman	0.37229	0.39873	2	2
Kendall	0.36000	0.36751	1	1
Median	0.37907	0.39881	3	3
Quadrant	0.37667	0.41745	4	4

Information from Table 34 to Table 36 displayed the results of contaminated normal data when rho = 0.3, outlier level = 10% and sample sizes were 10, 30 and 50 respectively. Also, it provides information about the performances of Pearson, Spearman, Kendall, Median and Quadrant correlation methods under Absolute Bias, Root Mean Square Error (RMSE).

Table 37: The Result of Absolute Bias and RMSE for rho = 0.9, level = 1%, n = 10

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.74338	0.86450	4	4
Spearman	0.62727	0.71983	1	1
Kendall	0.70000	0.73577	3	2
Median	0.66303	0.74900	2	3
Quadrant	0.966067	0.98489	5	5

Table 38: The Result of Absolute Bias and RMSE for rho = 0.9, level = 1%, n = 30

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.86558	0.87082	4	4
Spearman	0.83946	0.84729	1	1
Kendall	0.84713	0.86087	2	2
Median	0.85914	0.86335	3	3
Quadrant	0.94000	0.94227	5	5

Table 39: The Result of Absolute Bias and RMSE for rho = 0.9, level = 1%, n = 50

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.99345	0.99513	5	5
Spearman	0.94305	0.95738	1	1
Kendall	0.97537	0.97590	2	2
Median	0.98316	0.98477	3	3
Quadrant	0.99333	0.99405	4	4

Table 37 to Table 39 give information of contaminated normal data when rho = 0.9, outlier level = 1% and sample sizes of 10, 30 and 50 respectively. The performances of Pearson, Spearman, Kendall and Quadrant correlation methods in terms of Bias, Root Mean Square Error (RMSE) were also presented.

Table 40: The Result of Absolute Bias and RMSE for rho = 0.9, level = 5%, n = 10

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.85039	0.85848	5	5
Spearman	0.67172	0.67610	3	3
Kendall	0.68519	0.69029	4	4
Median	0.38001	0.56806	1	1
Quadrant	0.56667	0.59722	2	2

Table 41: The Result of Absolute Bias and RMSE for rho = 0.9, level = 5%, n = 30

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.85475	0.87899	5	5
Spearman	0.78469	0.79133	3	3
Kendall	0.82414	0.82731	4	4
Median	0.75726	0.77884	2	2
Quadrant	0.74444	0.76279	1	1

Table 42: The Result of Absolute Bias and RMSE for rho = 0.9, level = 5%, n = 50

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	1.00789	1.02137	5	5
Spearman	0.90335	0.91171	2	2
Kendall	0.90136	0.90568	1	1
Median	0.95098	0.95162	3	3
Quadrant	0.99333	0.99405	4	4

Details from Table 40 to Table 42 showed the results of contaminated normal data for rho = 0.9, outlier level = 5% and sample sizes of 10, 30 and 50 respectively. It also gave information about the level of performances of Pearson, Spearman, Kendall, Median and Quadrant correlation estimators under Absolute Bias, Root Mean Square Error (RMSE).

Table 43: The Result of Absolute Bias and RMSE for rho = 0.9, level = 10%, n = 10

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	1.18661	1.24536	5	5
Spearman	0.93030	0.94228	1	1
Kendall	1.12222	1.16892	3	3
Median	1.18743	1.23337	4	4
Quadrant	1.10000	1.10000	2	2

Table 44: The Result of Absolute Bias and RMSE for rho = 0.9, level = 10%, n = 30

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.98214	0.99832	5	5
Spearman	0.86745	0.87261	1	1
Kendall	0.89617	0.89898	2	2
Median	0.90229	0.90363	3	3
Quadrant	0.92222	0.92436	4	4

Table 45: The Result of Absolute Bias and RMSE for rho = 0.9, level = 10%, n = 50

Method	Bias	RMSE	Ranking of Bias (low to high)	Ranking of RMSE (low to high)
Pearson	0.97917	0.99843	5	5
Spearman	0.90706	0.90822	1	1
Kendall	0.91007	0.91184	2	2
Median	0.91247	0.91266	3	3
Quadrant	0.94000	0.94227	4	4

Table 43 to Table 44 contained results of contaminated normal data for fixed rho = 0.9, outlier level 10% and sample sizes (n) were 10, 30 and 50 respectively. The performances of Pearson, Spearman, Kendall, Median and Quadrant correlation methods under absolute bias and RMSE were also presented.

CONCLUSION

In this study, five correlation methods were compared for efficiency under two properties: Absolute Bias and Root Mean Square Error (RMSE), for varied sample sizes and outlier levels, under non-contaminated and contaminated normal data.

From the findings, it can be concluded that under the condition of non-contaminated normal data, Pearson is the best (because it never fell in 3rd position in all categories), under both low and high correlation levels. Also, Quadrant method performed the least among other methods.

On the other hand, under contaminated normal data, when outlier level is small or large, at all sample sizes and the rho is fixed at low or high, the Spearman method is preferred, followed by Kendall Tau. Alternatively, Pearson method is the least performed method. the spearman showed robustness in every category of simulation (because never fell in 4th position in all categories), therefore, the spearman method can be said to be the overall highest performer.

In line with the findings and conclusion of this study, it is recommended that Pearson method is appropriate to adopt when data is non-contaminated. While, Spearman method followed by Kendall should employ for contaminated data.

REFERENCES

[1]. Abdullah, S. (2015). Robust Correlation Procedure via *Sn* Estimator. Journal of Telecommunication, Electronic and Computer Engineering, Vol. 10 No. 1-10

[2]. Anscombe, F. (1973). Graphs in Statistical Analysis. Am Statistician 27 :17-21.

[3]. Blomqvist, N. (1950). On a Measure of Dependence between Two Random Variables. The Annals of Mathematical Statistics, 21, 593-600.

[4]. Chok, S. (2010). Pearson’s Versus Spearman’s and Kendall’s Correlation Coefficient for Continuous Data. Master’s Thesis, University of Pittsburgh, Pittsburgh.

[5]. Coblick ,W. (1998). Studies in the History of Statistics Method, London: Arno Press

[6]. Fowler, R. (1987). Power and Robustness in Product-Moment Correlation. Applied Psychological Measurement, 11:419-428.

[7]. Galton, F.(1889). Natural and Inheritance. London and New York, Macmillan, Vol. 13, pp 266- 267.

[8]. Genest, C. (2003). On Blest’s Measure of Rank Correlation. The Canadian Journal of Statistics, Vol. 31, No 1,1-8.

[9]. Gupta, S.C (2011).*Fundamentals of statistics*. Mumbai: Himalaya Publishing House.

[10]. Keiser, C.(2010). Analysis of Steam Formation and Migration in Firefighters’ Protective Clothing Using X-Ray Radiography. International Journal of Occupation 16(2): 217-229.

[11]. Kendall, M. (1938). A new measure of rank correlation. Biometrika, 30, pp. 81-93.

[12]. Kozak, M. (2008). Correlation Coefficient and the Fallacy of Statistical HypothesisTesting. Curriculum of Science, 95(9): 1121-1122.

[13]. Leuven, K. (2012). Robustness versus efficiency for nonparametric correlation measures. Economics and Leuven Statistics Research Centre, Leuven, Belgium.

[14]. Onwuegbuzie, A. (1999). Uses and Misuses of the Correlation Coefficient. Paper Presented at the Annual Meeting of the Mid-South Educational Research Association, lahti, Finland.

[15]. Osborne, J. and Overbay, A. (2004). The Power of Outliers. Practical Assessment, Research & Evaluation, 9(6).

[16]. Peng et al. (2012). Robust Rank Correlation Based Screening. Institute of Mathematical Statistics, Vol. 40, No 3, 1846-1877.

[17]. Shafiullah, A. and Khan, J. (2012). A New Robust Correlation Estimator for Bivariate Data. Bangladesh Journal of Scientific Research, Vol. 24, No. 97-106.

[18]. Shevlyakov, G. and Smirnov, P. (2011). Robust Estimation of the Correlation Coefficient. Austrian Journal of Statistics, Vol. 40, No 1 & 2, 147-156.

[19]. Shevlyakov, G. and Vilchevsky, N. (2002). Minimax variance estimation of a correlation coefficient for epsilon-contaminated bivariate normal distributions. Statistics and Probability Letters, 57, 91-100.

[20]. Sinsomboonthong, J. (2016). Robust Estimators for the Correlation Measuer to Resist Outliers in Data. Journal Mathematical Fund Science, Vol. 48, No 3, 263-275

[21]. Torrico, J. and Janssens, M (2010). Rapid Assessment Methods of Resilience for Natural and Agricultural Systems. An Acad Bras Cienc 82: 1095-1105.

[22]. Tugran et al, (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.

[23]. Washington, D. (2010). On a Least Squares Adjustment of a Sampled Frequency Table, the Expected Marginal Totals are Known. Annals of Mathematical Statistics.11(4): 427-444.

[24]. Winter, J. and Gosling, S. (2016). Comparison the Pearson and Spearman Correlation Coefficient Across Distributions and Sample Sizes. Psychological Methods, 21(3), 273-290

Article DOI: 10.52589/AJMSS-LLNZXUOZ
DOI URL: https://doi.org/10.52589/AJMSS-LLNZXUOZ