Journal of Modern Applied Statistical Methods

Volume 16 | Issue 2 Article 13

December 2017

A Monte Carlo Study of the Effects of Variability and Outliers on the Linear Correlation Coefficient

Hussein Yousif Eledum
University of Tabuk, Saudi Arabia, heledum@yahoo.com

Journal of Modern Applied Statistical Methods
November 2017, Vol. 16, No. 2, 231-255.
doi: 10.22237/jmasm/1509495180

ISSN 1538 − 9472

A Monte Carlo Study of the Effects of Variability and Outliers on the Linear Correlation Coefficient

Hussein Yousif Eledum
University of Tabuk
Tabuk, Saudi Arabia

Monte Carlo simulations are used to investigate the effect of two factors, the amount of variability and an outlier, on the size of the Pearson correlation coefficient. Some simulation algorithms are developed, and two theorems for increasing or decreasing the amount of variability are suggested.

Keywords: association, correlation, Pearson correlation coefficient, outlier, sample size, variability, Monte Carlo

Introduction

A correlation describes the relationship between two variables. Although there are a number of different correlation statistics, the one that is used most often is the Pearson's correlation (*PC*) defined in terms of the population correlation rho, as

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

where $Cov(X,Y)$ is the correlation between X and Y, σ_X,σ_Y are the population standard deviations of X and Y respectively.

The corresponding sample correlation $r_{x,y}$ (or $\hat{\rho}_{x,y}$) defined by

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y})}{(n-1)s_X s_Y} \tag{2}$$

Dr. Eledum is an Associate Professor in the Department of Statistics. Email him at heledum@yahoo.com.

where, s_X and s_Y are the sample standard deviations of X and Y respectively. The term $\sum_{i=1}^{n}(x_i-\bar{x})(y-\bar{y})$ is the sample covariance. In terms of z-scores,

$$\rho_{x,y}=\frac{\sum_{i=1}^{N} z_X z_Y}{N} \tag{3}$$

where, z_X is the z-score of the X variable, calculate using the population μ_X, and standard deviation σ_X, z_Y is likewise the z-score of the Y variable, and N is the number of pairs of scores.

Many studies have been conducted to study factors affecting the size of the correlation coefficient. Goodwin & Leech (2006) discussed factors that affect the size of *PC*, and Bates et al. (1996) investigated the effects of variability as a function of sample size on the *PC* under assumption of perfect relationship between two variables. Osborne & Overbay (2004) used the NELS data set (a national longitudinal study of 8th Grade students attending 1,052 high schools across the United States) to see the effect of outliers on two different types of correlations. In the current study, a Monte Carlo simulation will be used to investigate the effects of variability and outliers on the size of *PC*. In order to generate such data some algorithms have been developed, and two theorems are suggested to increase (or decrease) the amount of variability.

Variability

Variability refers to how spread out a set of data is. The four main measures to describe variability in a data set are: range, interquartile range, variance, and standard deviation. Conceptually, the Pearson Correlation *PC* of equation 2, is the ratio of the variation shared by X and Y to the variation of X and Y separately. That is:

$$r_{x,y}=\frac{\text{shared variability of } X \text{ and } Y}{\text{separate variability of } X \text{ and } Y} \tag{4}$$

When there is a perfect linear relationship, every change in the X variable is accompanied by a corresponding change in the Y variable. In this case, all variation in X is shared with Y, so $r_{x,y} = 1$. At the other extreme, when there is no linear relationship between X and Y, then the numerator is zero, so $r_{x,y} = 0$. So, equation 4 indicates that definitely variability influence the size of *PC*. Looking at

equation 4 we observe that increase or decrease in variability of single variable *X* or *Y* increases or decreases the shared variability (numerator) and variability of *X* or *Y* (part of denominator). Also, increase or decrease in variability of both variables *X* and *Y* increases or decreases the shared variability (numerator) and separate variability of *X* and *Y* (denominator). Therefore, the size of *PC* increases if the nominator is greater (or decreases if less) than the denominator, and this depends only on the data set, sample size, and the amount of variability in *X*, *Y*, or both.

Glass & Hopkins (1996) noted the value of the correlation coefficient PC will be greater if there is more variability among the observations than if there is less variability. Peers (2006) mentioned a good sample design will minimize the amount of variability in observations. The reduction in variability of a variable has the effect of reducing the correlation a variable has with other variables. The simple correlation is impacted when the variances of two measures are different, such as might occur with a restricted range.

In terms of restriction of range, there are procedures available for the estimation of the correlation for the entire group from the correlation obtained with the selected group (Glass & Hopkins, 1996; Gulliksen, 1950; Nunnally & Bernstein, 1994; Thorndike, 1982). However, the equation used to estimate the unrestricted correlation requires knowledge of the standard deviations of *X* and *Y* for the entire group and also requires several assumptions that are rarely tenable in practical situations (Crocker & Algina, 1986). Furthermore, the obtained estimates are often imprecise unless the sample size *n* is very large (Gullickson & Hopkins, 1976; Linn, 1983). A way to increase or decrease variability is to concomitantly incease or decrease the range. The following two theorems were developed to reduce the variability in term of variance using the idea of reduction range of data set.

Theorem 1

Suppose $x_1, x_2, \ldots, x_n$ are n real positive numbers with mean $\bar{x}$ and variance s_X^2, such that $x_1, x_2, \ldots, x_{n-1} < x_n$, if x_n substituted by $\bar{x}$, let $\bar{x}^*$, s_X^{*2} be mean and variance for new data set respectively, then

(a) $\bar{x}^* < \bar{x}$ (b) $s_X^{*2} < s_X^2$

Proof.

Proof of part (a). First, prove that $\bar{x} < x_n$. The mean of original data is defined by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad (*)$$

According to above formula, $\bar{x} = x_n$ if and only if $x_1 = x_2 = \cdots = x_n$, but $x_1, x_2, \ldots, x_{n-1} < x_n$, therefore, $\bar{x} < x_n$.

Substituting x_n by $\bar{x}$ in formula *

$$\bar{x}^* = \frac{x_1 + x_2 + \cdots + \bar{x}}{n}$$

since $\bar{x} < x_n$, then $\bar{x}^* < \bar{x}$.

Proof of part (b). The sample variance of the original data is defined by

$$s_X^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

Replacing x_n by $\bar{x}$, obtain

$$s_X^{*2} = \frac{(x_1 - \bar{x}^*)^2 + (x_2 - \bar{x}^*)^2 + \cdots + (\bar{x} - \bar{x}^*)^2}{n-1}$$

Suppose $x_1, x_2, \ldots, x_{n-1} < \bar{x}$, because $\bar{x}^* < \bar{x}$ then,

$$(x_i - \bar{x})^2 > (x_i - \bar{x}^*)^2 \ \forall\ i = 1, 2, \ldots, n-1$$

then

$$\sum_{i=1}^{n-1} (x_i - \bar{x})^2 > \sum_{i=1}^{n-1} (x_i - \bar{x}^*)^2 \qquad (a1)$$

Let $w = \frac{1}{n}\sum_{i=1}^{n-1} x_i$ then,

$$\bar{x} = w + \frac{x_n}{n} \tag{b1}$$

$$\bar{x}^* = w + \frac{\bar{x}}{n} \tag{b2}$$

Adding -1 times equation b2 to equation b1, $\bar{x} - \bar{x}^* = \frac{x_n - \bar{x}}{n}$, since $n > 1$, therefore, $x_n - \bar{x} > \bar{x} - \bar{x}^*$ implies

$$\left(x_n - \bar{x}\right)^2 > \left(\bar{x} - \bar{x}^*\right)^2 \tag{a2}$$

Combining two inequalities of a1 and a2

$$\sum_{i=1}^{n-1}\left(x_i - \bar{x}\right)^2 + \left(x_n - \bar{x}\right)^2 > \sum_{i=1}^{n-1}\left(x_i - \bar{x}^*\right)^2 + \left(\bar{x} - \bar{x}^*\right)^2$$

Dividing each side of above inequality by $n - 1$ to obtain

$$\frac{\sum_{i=1}^{n-1}\left(x_i - \bar{x}\right)^2 + \left(x_n - \bar{x}\right)^2}{n-1} > \frac{\sum_{i=1}^{n-1}\left(x_i - \bar{x}^*\right)^2 + \left(\bar{x} - \bar{x}^*\right)^2}{n-1} \text{ or } s_X^2 > s_X^{*2}$$

Therefore, $s_X^{*2} < s_X^2$.

Corollary

Suppose $x_1, x_2, \ldots, x_n$ are n real positive numbers with mean $\bar{x}$ and variance s_X^2, such that $x_1, x_2, \ldots, x_{n-1} > x_n$, if x_n substituted by $\bar{x}$, let $\bar{x}^*$, s_X^{*2} be mean and variance for new data set respectively, then

(a) $\bar{x}^* > \bar{x}$ (b) $s_X^{*2} > s_X^2$

Proof. Follow the same steps used for the proof of Theorem 1 above.

Theorem 2

Suppose $x_1, x_2, \ldots, x_n$ are n real positive numbers with mean $\bar{x}$ and variance s_X^2, such that $x_1, x_2, \ldots, x_{n-2} < x_{n-1}, x_n$, and $x_{n-1} = x_n$. If x_n was substituted by $\bar{x}$, to get new data set 1 $x_1, x_2, \ldots, x_{n-1}, \bar{x}$ with mean $\bar{x}^*$ and variance s_X^{*2} respectively. Suppose x_{n-1} in a new data set 1 substituted by $\bar{x}^*$, let $\bar{x}^{**}$, s_X^{**2} be mean and variance for new data set 2 respectively, then

$$\text{(a) } \bar{x}^{**} < \bar{x}^* \qquad \text{(b) } s_X^{**2} < s_X^{*2}$$

Proof.

Proof of part (a). First, prove that $\bar{x}^* \le x_{n-1}$. The mean of the new data set 1 is defined by

$$\bar{x}^* = \frac{x_1 + x_2 + \cdots + \bar{x} + x_{n-1}}{n} \qquad (**)$$

Because $x_{n-1} = x_n$ and $\bar{x} < x_n$ (Theorem 1) then $\bar{x} < x_{n-1}$.

According to formula **, $\bar{x}^* = x_{n-1}$ if and only if $x_1 = x_2 = \cdots = \bar{x} = x_{n-1}$, but $x_1, x_2, \cdots, x_{n-2}, \bar{x} < x_{n-1}$, therefore, $\bar{x}^* < x_{n-1}$. Because $\bar{x}^* < x_{n-1}$, then $\bar{x}^{**} < \bar{x}^*$.

Proof of part (b). The variances of new data set 1 are defined by

$$s_X^{*2} = \frac{\left(x_1 - \bar{x}^*\right)^2 + \left(x_2 - \bar{x}^*\right)^2 + \cdots + \left(x_{n-2} - \bar{x}^*\right)^2 + \left(x_{n-1} - \bar{x}^*\right)^2 + \left(\bar{x} - \bar{x}^*\right)^2}{n-1}$$

Replacing x_{n-1} by $\bar{x}^*$,

$$s_X^{**2} = \frac{\left(x_1 - \bar{x}^{**}\right)^2 + \left(x_2 - \bar{x}^{**}\right)^2 + \cdots + \left(x_{n-2} - \bar{x}^{**}\right)^2 + \left(x_{n-1} - \bar{x}^{**}\right)^2 + \left(\bar{x} - \bar{x}^{**}\right)^2}{n-1}$$

Suppose $x_1, x_2, \cdots, x_{n-2}, \bar{x} < \bar{x}^*$, since $\bar{x}^{**} < \bar{x}^*$ then,

$$\left(x_i - \bar{x}^*\right)^2 > \left(x_i - \bar{x}^{**}\right)^2 \ \ \forall \ \ i = 1, 2, \ldots, n-2 \text{ and } \left(\bar{x} - \bar{x}^*\right)^2 > \left(\bar{x} - \bar{x}^{**}\right)^2$$

Therefore

$$\sum_{i=1}^{n-2}\left(x_i-\bar{x}^{*}\right)^2+\left(\bar{x}-\bar{x}^{*}\right)^2>\sum_{i=1}^{n-2}\left(x_i-\bar{x}^{**}\right)^2+\left(\bar{x}-\bar{x}^{**}\right)^2 \qquad \text{(c1)}$$

Now let $w=\frac{1}{n}\sum_{i=1}^{n-2}x_i+\frac{\bar{x}}{n}$ then,

$$\bar{x}^{*}=w+\frac{x_{n-1}}{n} \qquad \text{(d1)}$$

$$\bar{x}^{**}=w+\frac{\bar{x}^{*}}{n} \qquad \text{(d2)}$$

Adding -1 times equation d2 to equation d1, $\bar{x}^{*}-\bar{x}^{**}=\frac{x_{n-1}-\bar{x}^{*}}{n}$, since $n>1$, therefore, $x_{n-1}-\bar{x}^{*}>\bar{x}^{*}-\bar{x}^{**}$ implies that

$$\left(x_{n-1}-\bar{x}^{*}\right)^2>\left(\bar{x}^{*}-\bar{x}^{**}\right)^2 \qquad \text{(c2)}$$

Combining two inequalities of c1 and c2 obtain

$$\sum_{i=1}^{n-2}\left(x_i-\bar{x}^{*}\right)^2+\left(\bar{x}-\bar{x}^{*}\right)^2+\left(x_{n-1}-\bar{x}^{*}\right)^2>\sum_{i=1}^{n-2}\left(x_i-\bar{x}^{**}\right)^2+\left(\bar{x}-\bar{x}^{**}\right)^2+\left(\bar{x}^{*}-\bar{x}^{**}\right)^2$$

Dividing both sides of above inequality by $n-1$ to obtain

$$\frac{\left\{\sum_{i=1}^{n-2}\left(x_i-\bar{x}^{*}\right)^2+\left(\bar{x}-\bar{x}^{*}\right)^2+\left(x_{n-1}-\bar{x}^{*}\right)^2\right\}}{n-1}>\frac{\left\{\sum_{i=1}^{n-2}\left(x_i-\bar{x}^{**}\right)^2+\left(\bar{x}-\bar{x}^{**}\right)^2+\left(\bar{x}^{*}-\bar{x}^{**}\right)^2\right\}}{n-1} \text{ or } s_X^{*2}>s_X^{**2}$$

Therefore, $s_X^{**2}<s_X^{*2}$.

Outliers

Outliers can be defined as a data point far outside the norm for a variable or population (see, e.g. Jarrell, 1994; Rasmussen, 1988; Stevens, 1984). Hawkins (1980) described outlier as an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism. Outliers have also been defined as values that are dubious in the eyes of the researcher (Dixon, 1950) and contaminants (Wainer, 1976). Generally, outliers can be defined as a score, case, or subject that falls outside the range of the rest of the scores, cases, or subjects.

Outlier can also be defined in terms of distributions rather than numerical distance between observations. Therefore, distribution of order statistics from independent non-identical random variables are closely related with the outlier models. Barnett and Lewis (1994) considered the single-outlier model. Balakrishnan (2007) focused on the multiple-outlier model. He presented many results on order statistics from multiple-outlier models and illustrated their use in robustness studies.

Balakrishnan (1988) derived recurrence relations among moments of Order Statistics from two related Outlier models. Balakrishnan (1994a, 1994b) obtained recurrence relations for the single and product moments from non-identical exponential distribution and its right truncated. Balakrishnan and Balsubramanian (1995) gave recurrence relations for moments from non-identical power function distribution. Childs and Balakrishnan (1998) obtained recurrence relations for moments from non-identical Pareto and truncated Pareto distribution. Childs (2001) gave recurrence relations for the single and product moments from non-identical right truncated Lomax distribution. Moshref (2000) established recurrence relations for moments from non-identical generalized power function. Mahmoud et al. (2005) derived order statistics from non-identical generalized Pareto random variables. Recurrence relations for moments for Logistic from non-identical random variables have obtained by Childs and Balakrishnan (2006).

Outliers are often caused by human error, such as errors in data collection, recording, or entry. Sampling errors is another reason for outliers to be occurred, it is possible that a few members of a sample were inadvertently drawn from a different population than the rest of the sample (Osborne & Overbay, 2004). Outliers can also be caused by research methodology, or by incorrect assumptions about the distribution of the data (Iglewicz & Hoaglin, 1993). Barnett and Lewis (1994) explained not all outliers are illegitimate contaminants, and not all illegitimate scores show up as outliers. Generally, outliers can be classified into

two major categories, those due to errors in the data, and those from the inherent variability of the data.

The presence of outlier can result in an increase or decrease in the size of *PC*, depending on the location of the outlier (Glass & Hopkins, 1996). Stockburger (2013) demonstrated outlier that falls near where the regression line would normally fall would necessarily increase the size of the correlation coefficient. An outlier that falls some distance away from the original regression line would decrease the size of the correlation coefficient. They also illustrated that smaller the sample size, the greater the effect of the outlier, and at some point the outlier will have little or no effect on the size of the correlation coefficient.

There are various methods of outlier detection; one simple way is to examine the scatter diagram, another method is to use the rules of thumb (data points three or more standard deviations from the mean, or 1.5 IQR criterion). Some researchers prefer visual inspection of the data. Lornez (1987) argued outlier detection is a special case of the examination of data for influential data points.

If there exists an outlier on the dataset, first check for human error (errors in data collection, recording, or entry). If there are no justifications for categorizing the datum an outlier, it should not be removed from the analysis.

Monte Carlo Simulation

A computer program using R Version 3.3.3 was developed as follows.

Algorithm 1

Step 1.1. *Population 1 of size 1,000,000*: Generating random variable X follows normal distribution with a mean μ_X and a standard deviation of σ_X.

Step 1.2. *Population 2 of size 1,000,000*: Generating another random variable Y follows normal distribution with a mean μ_Y and standard deviation σ_Y, correlated with X with a particular population ρ.

Step 1.3. *Sample of size n*: Selecting sample of size n at randomly from each population. Then Algorithm 2 (or 3) is executed.

Step 1.4. *Replication*: Procedures of Step 1.3 were repeated 100,000 times, and the overall average of these repetitions is computed.

To examine the effect of this two factors on the size of *PC*, some different values of ρ were set, that is, 0.002 (weak correlation), 0.5 (moderate correlation) and 0.99 (strong correlation). Furthermore, sample of sizes 20, 60, 120, and 360 have been determined. Equation 1 is used to create the variance covariance matrix $\boldsymbol{\Sigma}$ for the univariate normal distribution of the two variables *X* and *Y*,

$$\Sigma = \begin{bmatrix} 1 & \operatorname{cov}(X,Y) \\ \operatorname{cov}(X,Y) & 1 \end{bmatrix} \text{ where } \operatorname{cov}(X,Y) = \rho_{x,y} = \sigma_X \sigma_Y$$

Variability

To illustrate the relationship between variability and the size of *PC*, follow the two steps of Algorithm 1 by setting $\mu_X = 10$, $\mu_Y = 20$, $\sigma_X = \sigma_Y = 1$, for the values of ρ we seleted only the high correlation i.e. $\rho_{x,y} = 0.99$, then developed Algorithm 2.

Algorithm 2

Step 2.1. After generating *N* pairs of data points (*X*,*Y*), with population correlation rho, the data were arranged in ascending order on *X* and *Y*. *PC*s for new variables were calculated and stored.

Step 2.2. To conduct the effect of variability on the size of *PC*, reduce the amount of variability using Theorem 2 after some modifications. Reduction of variability included

1. Both variables *X* and *Y* gradually by deleting the highest 5%, 10%, and 20% values each time.

2. Single variable *Y* by deleting the highest 5%, 10%, and 20% values each time.

To avoid decrease of the sample size, substitute the deleted values by the averages of *X* and *Y* for (1) and the average of *Y* for (2).

Compiled in Table 1 are the variance of *X*, variance of *Y*, and size of *PC* for original samples and samples (new samples) after (1) of Step 2.2 applied, for each

sample size. Compiled in Table 2 are the variance of *Y*, size of *PC* for original samples , and samples (new samples) after (2) of Step 2.2 applied.

Note that

- The value between two brackets represents a percentage of reduction in variance of *X*, variance of *Y*, and size of *PC* for new samples with respect to original samples.

- Newsample1, Newsample2, and Newsample3 represent sample (new samples) after the highest 5%, 10%, and 20% values of original sample have been deleted and substituted by the average of specific variable respectively.

Table 1. Var(*X*), Var(*Y*), and size of *PC* for original and new samples for each sample size.

n	Data	Var(*X*)	Var(*Y*)	*PC*
20	Original sample	0.999802	0.9998876	0.9940756
	Newsample1	0.7945104 (20.53323)	0.7946805 (20.52302)	0.9931182 (0.09631254)
	Newsample2	0.6647697 (33.50986)	0.6650363 (33.4889)	0.9923499 (0.1736026)
	Newsample3	0.4984001 (50.1501200)	0.4987495 (50.11945)	0.9910805 (0.301297)
60	Original sample	0.9983744	0.9971149	0.9968837
	Newsample1	0.7770588 (22.1676)	0.7764702 (22.12832)	0.996571 (0.04099317)
	Newsample2	0.6489883 (34.9955)	0.648791 (34.93318)	0.9961512 (0.07347576)
	Newsample3	0.4879877 (51.12178)	0.4878847 (51.07036)	0.9955679 (0.13199)
120	Original sample	1.004146	1.003993	0.9981221
	Newsample1	0.7768069 (22.64006)	0.7767548 (22.63348)	0.9979473 (0.01751774)
	Newsample2	0.6497357 (35.29471)	0.6496027 (35.29811)	0.9977732 (0.03496057)
	Newsample3	0.48919 (51.283)	0.4891632 (51.27825)	0.9974267 (0.06967345)
360	Original sample	0.9963796	0.9957264	0.9992268
	Newsample1	0.7699424 (22.726)	0.7690821 (22.7617)	0.9992059 (0.002093478)
	Newsample2	0.6445929 (35.3065)	0.644101 (35.31345)	0.9991336 (0.009329307)
	Newsample3	0.4861082 (51.21254)	0.4853523 (51.25645)	0.9989874 (0.0239634)

Table 2. Var(*Y*) and size of *PC* for original and new samples for each sample size

	Original Sample		**Newsample1**	
n	Var(*Y*)	*PC*	Var(*Y*)	*PC*
20	0.9998876	0.9940756	0.7946805	0.8954083
			(20.52302)	(9.925539)
60	0.9971149	0.9968837	0.7764702	0.8903554
			(22.12832)	(10.68614)
120	1.003993	0.9981221	0.7767548	0.8894803
			(22.63348)	(10.88462)
360	0.9957264	0.9992268	0.7690821	0.8902479
			(22.7617)	(10.90632)

	Newsample2		**Newsample3**	
n	Var(*Y*)	*PC*	Var(*Y*)	*PC*
20	0.6650363	0.8442923	0.4987495	0.8095944
	(33.4889)	(15.0676)	(50.11945)	(18.5581)
60	0.648791	0.8403698	0.4878847	0.8077327
	(34.93318)	(15.70031)	(51.07036)	(18.97423)
120	0.6496027	0.8404212	0.4891632	0.8082734
	(35.29811)	(15.79976)	(51.27825)	(19.02059)
360	0.644101	0.8415755	0.4853523	0.8097937
	(35.31345)	(15.7733)	(51.25645)	(18.95797)

For $n = 20$, a reduction of 20.5%, 33,5%, and 50.1% in the variances of both variables *X* and *Y*, results in reduction of 0.0963%, 0.1736%, and 0.3013% in the size of *PC* respectively. The same reductions in the variance of *Y* led to reduction of 9.925%, 15.068%, and 18.558% in the size of *PC*. When $n = 60$, a reduction of 22.2%, 34.9%, and 51.1% in the variances of both *X* and *Y* yields 0.0409%, 0.0735%, and 0.1319% reductions in the size of *PC* respectively. The same reductions in the variance of *Y* results in 10.686%, 15.700%, and 18.974% reductions in the size of *PC*.

The same reductions as $n = 20$ and 60 in the variances of both variables, yield a reduction of 0.0175%, 0.0349%, and 0.0697% for $n = 120$, and 0.0021%, 0.0093%, and 0.02396% for $n = 360$ in the size of *PC* respectively. Whereas, reductions in *Y* alone for these two sample sizes follow the same pattern of $n = 20$ and 60.

Accordingly, the following conclusions are advanced:

1. As the percentage of deleting highest values from original sample increases, the percentage of reduction in *PC* increases for all sample sizes. This can be seen also from Figure 1, and it means that as the amount of variability increases the size of *PC* decreases.

2. As the sample size increases the percentage of reduction in the size of *PC* decreases, also the effect of reduction in variances of two variables *X* and *Y* on the size of the *PC* decreases as the sample size increases (see Figure 1).

3. The effect of reduction in variance of *Y* alone on the size of *PC* is not affected by the sample size.

4. A reduction in the variance of *Y* alone has strong effect on the size of the *PC* than a reduction in the variances of two variables *X* and *Y*.

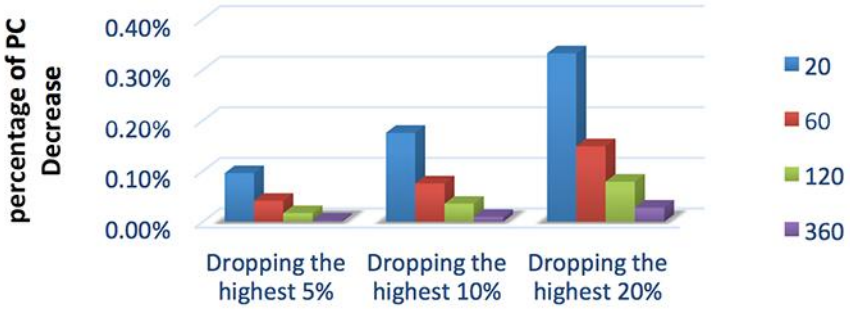

Figure 1. Percentage of reduction in PC of new samples for each sample size

The effect of an Outlier

To study the impact of an outlier on the size of PC, the two steps of Algorithm 1 have been followed after setting $\mu_X = 10$, $\mu_Y = 10$, $\sigma_X = 1$, and $\sigma_Y = 1$, and some values of ρ have been set, that is, 0.002 (weak correlation), 0.05 (moderate correlation) and 0.99 (strong correlation), then Algorithm 3 below has been designed.

Algorithm 3

Step 3.1. *Creating an outlier*: Add a single observation out of the samples ranges that were selected in Step 1.3; this observation represents an outlier. This is done for all samples of each variable *X* and *Y* separately, and for both variables at the same time. Take into account the position of this observation from regression line and other observations. Then compute the size of *PC* between *X* and *Y* for each case. In Table 3, outliers and their outlier distant from other observations of two variables *X* and *Y* are given.

Table 3. The created outliers for each variable X and Y.

X / ***Y***	$\mu_X + 0\sigma_X$	$\mu_X + 4\sigma_X$	$\mu_X + 6\sigma_X$	$\mu_X + 8\sigma_X$	$\mu_X + 10\sigma_X$
$\mu_Y + 4\sigma_Y$	✓	✓			
$\mu_Y + 6\sigma_Y$	✓	✓	✓		
$\mu_Y + 8\sigma_Y$	✓	✓	✓	✓	
$\mu_Y + 10\sigma_Y$	✓	✓	✓	✓	✓

The check symbol (✓) in Table 3 above indicates this data point (outlier) is done; for example, the shaded cell with the check symbol implies that the created outlier is $\mu_X + 6\sigma_X$ for variable *X* and $\mu_Y + 4\sigma_Y$ for variable *Y*, where μ_X, μ_Y are the averages of two populations, and σ_X, σ_Y are standard deviations. Set $\mu_X = \mu_Y = 10$ and $\sigma_X = \sigma_Y = 1$, and therefore, the data point (x,y) corresponding the shaded cell is (16,14).

Complied in Tables 4 - 7 are an outlier and the size of *PC* for sample sizes 20, 60, 120 and 360 for each value of $\hat{\rho}$. The value between parentheses represents the percentages of increase in the size of *PC* after the existence of outlier.

Table 4. An outlier and size of *PC* for each value of $\hat{\rho}$ when *n* = 20.

$\hat{\rho}$	Outlier	10	14	16	18	20
0.000134	14	0.0004155 (207.961)	0.4540743 (336402.4)			
	16	0.0004376 (224.305)	0.5427325 (402104.6)	0.648574 (480540)		
	18	0.0004526 (235.482)	0.589405 (436692.4)	0.704355 (521879)	0.7649272 (566767)	
	20	0.0004652 (244.757)	0.6157849 (456241.8)	0.735883 (545243)	0.7991673 (592141)	0.834962 (618667)
0.517289	14	0.3832509 (-25.9117)	0.7396204 (42.97996)			
	16	0.3077681 (-40.5037)	0.7721365 (49.26581)	0.833112 (61.053)		
	18	0.251873 (-51.3089)	0.7771106 (50.22738)	0.855385 (65.359)	0.8886333 (71.7864)	
	20	0.2111501 (-59.1815)	0.7730817 (49.44853)	0.862465 (66.727)	0.9028568 (74.5361)	0.921919 (78.2211)
0.9989	14	0.7331424 (-26.6086)	0.999444 (0.049636)			
	16	0.5864489 (-41.2934)	0.9799945 (-1.897359)	0.999646 (0.06982)		
	18	0.4786875 (-52.0808)	0.946635 (-5.236816)	0.991416 (-0.75405)	0.9997641 (0.08168)	
	20	0.4005459 (-59.9032)	0.9147883 (-8.424848)	0.976297 (-2.2675)	0.9959278 (-0.30235)	0.999835 (0.08877)

Table 5. An outlier and size of *PC* for each value of $\hat{\rho}$ when $n = 60$.

$\hat{\rho}$	Outlier	10	14	16	18	20
0.0032	14	0.002748 (-15.028)	0.216529 (6595.58)			
	16	0.002391 (-26.071)	0.286851 (8770.12)	0.380909 (11678.6)		
	18	0.002057 (-36.378)	0.335072 (10261.2)	0.445228 (13667.5)	0.5209696 (16009.6)	
	20	0.002057 (-45.231)	0.367746 (11271.5)	0.488834 (15015.9)	0.5721052 (17590.9)	0.6286688 (19339.94)
0.5278	14	0.467915 (-11.345)	0.629935 (19.3529)			
	16	0.415922 (-21.196)	0.654849 (24.0732)	0.708196 (34.1809)		
	18	0.365841 (-30.685)	0.659175 (24.8929)	0.733548 (38.9843)	0.7745937 (46.7612)	
	20	0.322084 (-38.975)	0.653423 (23.8031)	0.743026 (40.7801)	0.7957244 (50.7648)	0.8254869 (56.4038)
0.9989	14	0.884845 (-11.425)	0.999202 (0.02223			
	16	0.785982 (-21.322)	0.982626 (-1.6371)	0.999371 (0.03919)		
	18	0.690923 (-30.837)	0.947056 (-5.1977)	0.989356 (-0.9633)	0.9995146 (0.05355)	
	20	0.607981 (-39.139)	0.906502 (-9.2572)	0.967956 (-3.1055)	0.9935535 (-0.54316)	0.9996243 (0.064532)

Table 6. An outlier and size of *PC* for each value of $\hat{\rho}$ when n = 120.

$\hat{\rho}$	Outlier	10	14	16	18	20
0.0067	14	0.006246 (-6.7973)	0.124657 (1759.99)			
	16	0.005812 (-13.266)	0.171539 (2459.53)	0.23751 (3443.88)		
	18	0.005336 (-20.380)	0.208615 (3012.74)	0.289602 (4221.14)	0.3538695 (5180.07)	
	20	0.004866 (-27.397)	0.237067 (3437.27)	0.329624 (4818.31)	0.4030966 (5914.59)	0.4596879 (6758.984)
0.498478	14	0.468111 (-6.0919)	0.558143 (11.9694)			
	16	0.437 (-12.333)	0.576238 (15.5995)	0.615193 (23.4142)		
	18	0.402374 (-19.279)	0.581287 (16.6123)	0.637398 (27.8688)	0.6739801 (35.2075)	
	20	0.368035 (-26.168)	0.577984 (15.9497)	0.647793 (29.9542)	0.6959424 (39.6134)	0.7274151 (45.92715)
0.999	14	0.937906 (-6.1159)	0.999124 (0.01188)			
	16	0.875407 (-12.372)	0.987769 (-1.1246)	0.999237 (0.02323)		
	18	0.805878 (-19.332)	0.960065 (-3.8979)	0.990903 (-0.8110)	0.9993538 (0.03492)	
	20	0.736952 (-26.231)	0.924295 (-7.4784)	0.971021 (-2.8012)	0.9934752 (-0.55351)	0.9994598 (0.045533)

Table 7. An outlier and size of *PC* for each value of $\hat{\rho}$ when n = 360.

$\hat{\rho}$	Outlier	10	14	16	18	20
0.000459	14	0.000409 (-10.815)	0.042961 (9255.71)			
	16	0.000379 (-17.301)	0.062564 (13524.7)	0.091253 (19772.4)		
	18	0.000348 (-24.137)	0.080469 (17424.1)	0.117436 (25474.3)	0.1511931 (32825.8)	
	20	0.000316 (-31.062)	0.096466 (20907.6)	0.140828 (30568.6)	0.1813409 (39391.1)	0.2175222 (47270.46)
0.499015	14	0.488268 (-2.1537)	0.520538 (4.31308)			
	16	0.475749 (-4.6623)	0.528039 (5.81613)	0.544929 (9.20103)		
	18	0.459742 (-7.8701)	0.530407 (6.29067)	0.556016 (11.4226)	0.5751654 (15.2601)	
	20	0.441361 (-11.554)	0.52852 (5.91259)	0.562015 (12.6249)	0.588533 (17.9389)	0.608581 (21.95636)
0.999	14	0.977556 (-2.1471)	0.999048 (0.00424)			
	16	0.952604 (-4.6448)	0.994232 (-0.4778)	0.999096 (0.00906)		
	18	0.92069 (-7.8393)	0.980909 (-1.8115)	0.994835 (-0.4174)	0.9991558 (0.01504)	
	20	0.884029 (-11.509)	0.96103 (-3.8013)	0.983257 (-1.5764)	0.9954962 (-0.35134)	0.9992218 (0.021644)

At data points (14,14), (16,16), (18,18), and (20,20), the size of *PC* increases for all sample sizes and all values of $\hat{\rho}$. Also, at this data points, the percentage of increase in the size of *PC* goes up as data points distant away from *X* and *Y* coordinates, that is, the percentage of increase at (20,20) greater than percentage of increase at the other data points. This can be seen in Figures 2a, 2b, and 2c. The reason is these data points follow the same pattern of the remainder of the data, in other words, some data points fall near the regression line and other fall in the same direction of the regression line if it is extended (see Figures 3a, 3b, and 3c).

At data points (x,10) where x = 14, 16, 18, 20, the size of *PC* decreases when n = 60, 120, and 360 for all values of $\hat{\rho}$, and when n = 20 for $\hat{\rho}$ = 0.000134, whereas *PC* increases when n = 20 for $\hat{\rho}$ = 0.517,0.998, because these data points lie at the bottom of the regression line top of the x coordinate

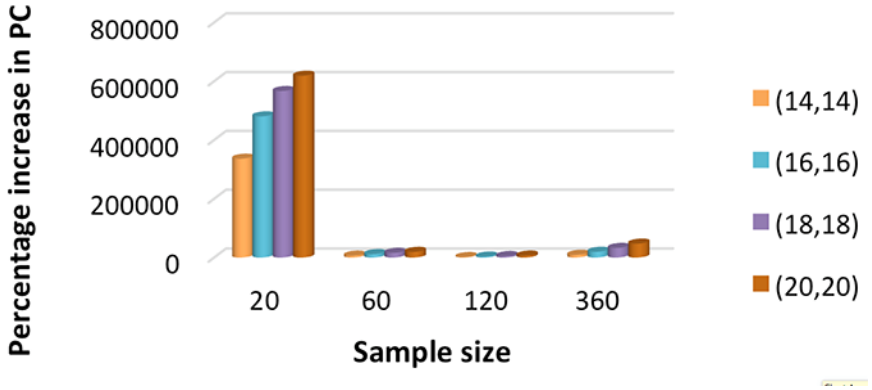

Figure 2a. Percentage of increase in *PC* for each sample size at $\hat{\rho}$ close to 0.

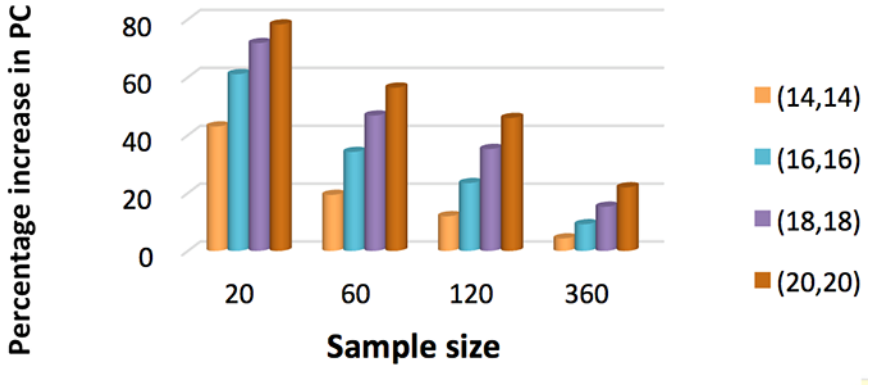

Figure 2b. Percentage of increase in *PC* for each sample size at $\hat{\rho}$ close to 0.5.

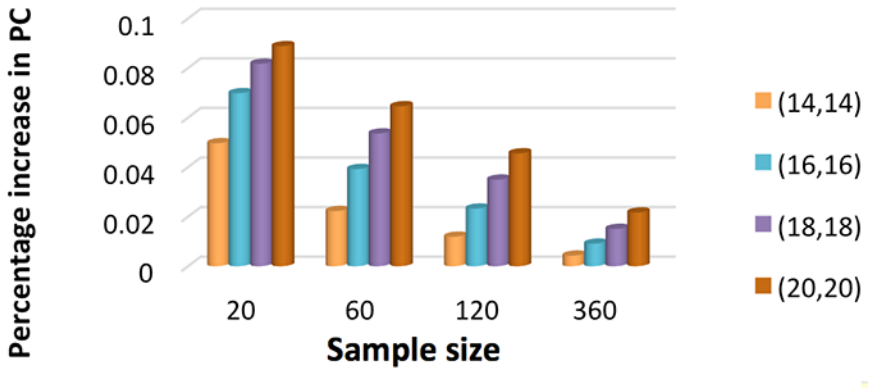

Figure 2c. Percentage of increase in *PC* for each sample size at $\hat{\rho}$ close to 1.

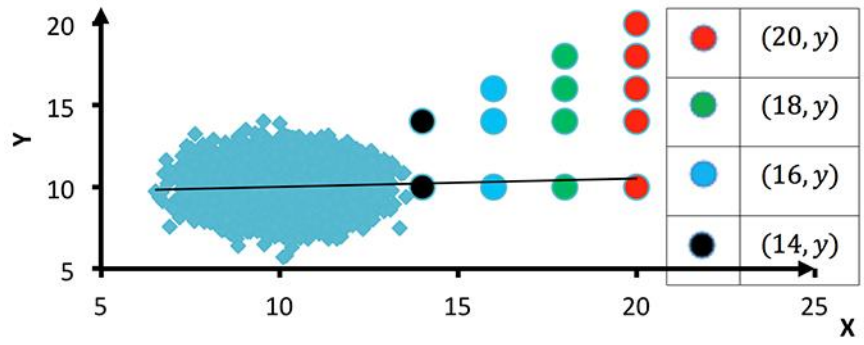

Figure 3a. The outliers on scatter diagram when ρ close to 0.

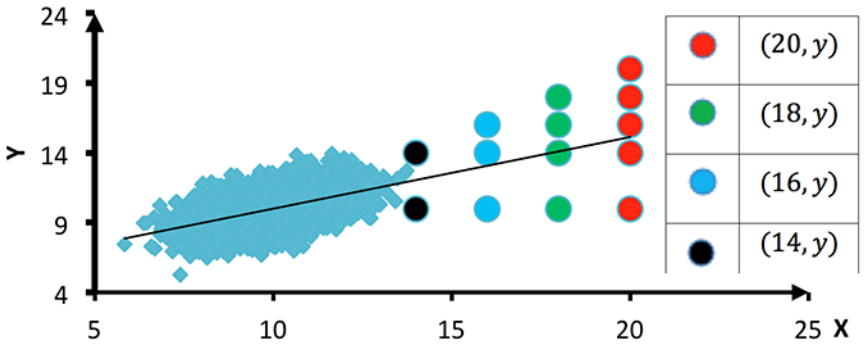

Figure 3b. The outliers on scatter diagram when ρ close to 0.5.

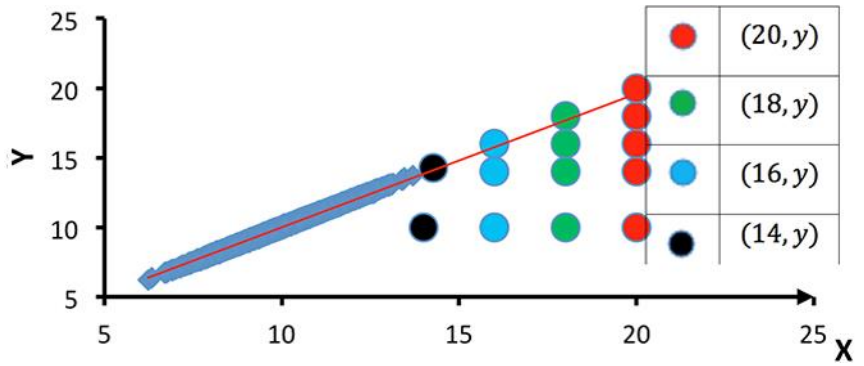

Figure 3c. The outliers on scatter diagram when ρ close to 1.

when $n = 60, 120, 360$, and top of the regression line when $n = 20$ for $\hat{\rho} = 0.517$, 0.998, and in the same direction of regression line at $n = 20$ for $\hat{\rho} = 0.000134$ (Figures 3a, 3b, and 3c). Moreover, the percentage of decrease in the size of *PC* at this data points goes down as the sample size increases ($n \neq 20$).

At data points (x,y) where $x \neq y$, the size of *PC* decreases for all sample sizes and $\hat{\rho}$ close to 1, whereas, for other two values of $\hat{\rho}$ the size of *PC* increases. This happens because the data points distant away from the rest of observations at the bottom of the regression line when $\hat{\rho}$ close to 1 (Figure 3c), and locates close to the rest of observations or in the same direction of the regression line when $\hat{\rho}$ close to zero and 0.5 (see Figures 3a and 3b).

The conclusions that may now be drawn are:

1. The existence of an outlier on data set might increase or decrease the size of *PC*, according to the position of this outlier from the rest of observations and the regression line.

2. The effect of an outlier on the size of *PC* decreases as the sample size increases.

3. Location of an outlier rather than its magnitude, determine the amount of its effect on the size of *PC*.

4. The effect of an outlier in the size of *PC* becomes more sensitive as the value of *PC* close to 1.

Conclusion

Several conclusions can be drawn from present study; the most important of them are: (a) As the amount of variability increases the size of *PC* decreases, (b) The effect of increase or decrease in the amount of variability in both variables on the size of *PC* decreases as the sample size increases, whereas the effect of increase or decrease in the amount of variability in single variable on the size of *PC* is not affected by the sample size, (c) a reduction in the variance of *Y* alone has a stronger effect on the size of the *PC* than a reduction in the variances of both variables *X* and *Y*, (d) everything else being equal, as the sample size increases the effect of an outlier on the size of *PC* decreases, and (e) the amount of effect of an outlier on the size of *PC* depends on factors including the amount of an outlier,

location of an outlier from the regression line or from the rest of observations, the size of *PC* itself, and the sample size.

References

Balakrishnan, N. (1988). Recurrence relations among moments of order statistics from two related outlier models. *Biometrical Journal, 30*(6), 741–746. doi: 10.1002/bimj.4710300619

Balakrishnan, N. (1994a). On order statistics from non-identical exponential random variables and some applications. *Computational Statistics and Data Analysis, 18*(2), 203 - 253. doi: 10.1016/0167-9473(94)90172-4

Balakrishnan, N. (1994b). On order statistics from non-identical right-truncated exponential random variables and some applications. *Communication in Statistics - Theory & Methods, 23*(12), 3373 - 3393. doi: 10.1080/03610929408831453

Balakrishnan, N. (2007). Permanents, order statistics, outliers, and robustness. *Revista Matemática Complutense, 20*(1), 7-107. doi: 10.5209/rev_rema.2007.v20.n1.16528

Balakrishnan, N. & Balasubramanian, K. (1995). Order statistics from non-identically power function random variables. *Communication in Statistics - Theory & Methods, 24*(6), 1443 - 1454. doi: 10.1080/03610929508831564

Barnett, V. & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.

Bates, B. T., Zhang, S., Dufek, J. S., Chen, F. C. (1996). The effects of sample and variability on the correlation coefficient. *Medicine and Science in Sports and Exercise, 28*(3), 386-391. doi: 10.1097/00005768-199603000-00015

Childs, A. & Balakrishnan, N. (1998). Generalized recurrence relations for moments of order statistics from non-identical Pareto and truncated Pareto random variables with applications to robustness. In N. Balakrishnan and C. R. Rao, Eds. *Handbook of Statistics: Order Statistics: Theory and Methods* (Volume 16). Amsterdam, Netherlands: Elsevier Science B. V. doi: 10.1016/s0169-7161(98)16017-0

Childs, A., Balakrishnan, N., & Moshref, M. (2001). Order statistics from non-identical right-truncated Lomax random variables with applications. *Statistical Papers, 42*(2), 187–206. doi: 10.1007/s003620100050

Childs, A. & Balakrishnan, N. (2006). Relations for order statistics from non-identical logistic random variables and assessment of the effect of multiple outliers on bias of linear estimators. *Journal of Statistics Planning and Inference, 136*(7), 2227-2253. doi: 10.1016/j.jspi.2005.08.026

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich.

Dixon, W. J. (1950). Analysis of extreme values. *Annals of Mathematical Statistics, 21*(4), 488-506. doi: 10.1214/aoms/1177729747

Glass, G. V. & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Goodwin, D. L. & Leech, L. N. (2006). Understanding correlation: factors that affect the size of r. *The Journal of Experimental Education, 74*(3), 251-266. doi: 10.3200/jexe.74.3.249-266

Gullickson, A. R. & Hopkins, K. D. (1976). Interval estimation of correlation coefficients corrected for restriction of range. Educational and *Psychological Measurement, 36*(1), 9–25. doi: 10.1177/001316447603600102

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley. doi: 10.1037/13240-000

Hawkins, D. M. (1980). *Identification of outliers*. London: Chapman and Hall. doi: 10.1007/978-94-015-3994-4

Iglewicz, B. & Hoaglin, D. C. (1993). *How to detect and handle outliers.* Milwaukee, WI.: ASQC Quality Press.

Jarrell, M. G. (1994). A comparison of two procedures, the Mahalanobis Distance and the Andrews-Pregibon Statistic, for identifying multivariate outliers. *Research in the Schools, 1*(1), 49-58.

Linn, R. L. (1983). Pearson selection formulas: implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement, 20*(1), 1–16. doi: 10.1111/j.1745-3984.1983.tb00185.x

Lornez, F. O. (1987). Teaching about influence in simple regression. *Teaching Sociology, 15*(2), 173-177. doi: 10.2307/1318032

Mahmoud, M., Moshref, M., & Sultan, K. (2005). Order statistics from non-identical generalized Pareto random variables with applications. *Journal of Applied Statistical Science, 14*(1), 85-98.

Moshref, M. (2000). Order statistics from non-identical generalized power function random variables and applications. *The Egyptian Statistical Journal, 44*(1), 99-111.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Osborne, J. W. & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation, 9*(6), 1-8.

Peers, I. S. (2006). *Statistical analysis for education and psychology researchers.* Washington, DC: The Falmer Press. doi: 10.4324/9780203985984

Rasmussen, J. L. (1988). Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D. *Multivariate Behavioral Research, 23*(2), 189-202. doi: 10.1207/s15327906mbr2302_4

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*(2), 334-344. doi: 10.1037/0033-2909.95.2.334

Stockburger, D. W. (2013). *Introductory Statistics: Concepts, Models, and Applications* (3rd Web Ed.). Springfield, MO: Author. Available at http://psychstat3.missouristate.edu/Documents/IntroBook3/sbk.htm

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.

Wainer, H. (1976). Robust statistics: A survey and some prescriptions. *Journal of Educational Statistics, 1*(4), 285-312. doi: 10.2307/1164985