Bias Correction in Estimation of the Population Correlation Coefficient

Juthaphorn Sinsomboonthong

ABSTRACT

An estimator of the population correlation coefficient of two variables for a bivariate normal distribution was proposed and evaluated using comparisons with the Pearson correlation coefficient and an estimator of Olkin and Pratt, conducted using a simulation study. It was found that for a small sample size of n=10, the absolute bias of the proposed estimator was less than those of the Pearson correlation coefficient and an estimator of Olkin and Pratt. In addition, the mean square errors of those estimators seemed to have no difference in each situation for this study.

Keywords: Pearson correlation coefficient, bivariate normal distribution, absolute bias, mean square error, estimator.

INTRODUCTION

The Pearson correlation coefficient is one of the most frequently used tools of researchers for correlation coefficient investigation as mentioned by Rodgers and Nicewander (1988) and Huson *et al.* (2007). Unfortunately, Neter *et al.* (1996) and Zimmerman *et al.* (2003) considered that it was a biased estimator for the population correlation coefficient (ρ). Furthermore, the bias decreased when the sample size increased and it was zero when the population correlation coefficient was zero or one. In addition, this conformed to the research of Sinsomboonthong (2011a). A conventional estimate of the Pearson correlation coefficient is likely to underestimate the population correlation coefficient (Gorsuch and Lehmann, 2010; Adolph and Hardin, 2007; Zimmerman *et al.*, 2003) because the distribution of this estimator is asymmetrical (Fisher, 1921).

Therefore, research activity to find a correction for bias in the estimation of the correlation coefficient has already been undertaken. Fisher (1915) published an approximately unbiased estimator of the population correlation coefficient in samples from an indefinitely large population. Later, Olkin and Pratt (1958) developed the Pearson correlation coefficient to decrease the amount of bias of this estimator for two variables having a bivariate normal distribution with equal variances. Then Zimmerman *et al.* (2003) demonstrated that the bias of the Pearson correlation coefficient was almost eliminated by Fisher's estimator and a related estimator proposed by Olkin and Pratt (1958).

In the present study, an estimator of the population correlation coefficient of two variables for a bivariate normal distribution was proposed and the jackknife method (Quenouille, 1949; Tukey, 1958) was applied for bias reduction

Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand.
E-mail: fscijps@ku.ac.th

Received date : 04/02/13 Accepted date : 08/05/13

(Efron and Tibshirani, 1993; Smith and Pontius, 2006; Sinsomboonthong, 2011a, b). Furthermore, comparisons of the absolute biases and mean square errors of three estimators—the proposed estimator, the Pearson correlation coefficient, and an estimator of Olkin and Pratt—were performed by a simulation study.

MATERIALS AND METHODS

This study proposed an estimator of the population correlation coefficient and applied the jackknife method for bias reduction of the Pearson correlation coefficient. In order to empirically evaluate the validity and reliability of the proposed estimator, a simulation study was conducted for 152 situations. Then, comparisons of the absolute bias and mean square error of the proposed estimator and two sample correlation coefficients—the Pearson correlation coefficient and the Olkin and Pratt estimator—were empirically performed.

Sample correlation coefficient

Let $(x_1, y_1),..., (x_n, y_n)$ be a random sample from a bivariate normal distribution with means μ_1, μ_1 variances σ_1^2, σ_2^2 and the population correlation coefficient ρ. It is well known that the maximum likelihood estimator of ρ, denoted by $\hat{\rho}_P$, is given by Equation 1

$$\hat{\rho}_P = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

and

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i \qquad (1)$$

(Neter *et al.*, 1996; Anderson, 2003). This estimator is often called the Pearson correlation coefficient. It is a biased estimator of ρ (unless ρ =0 or 1), which is usually small when the sample size is large (Neter *et al.*, 1996; Zimmerman *et al.*, 2003; Sinsomboonthong, 2011a). Later, Olkin and Pratt (1958) recommended the correction estimator in the form of $\hat{\rho}_{OP} = \hat{\rho}_P\left(1+\frac{1-\hat{\rho}_P^2}{2(n-3)}\right)$ as a more nearly unbiased estimator of ρ. Zimmerman *et al.* (2003) using simulation demonstrated that the bias of $\hat{\rho}_{OP}$ was less than that of $\hat{\rho}_P$ for a small sample size.

Proposed estimator

This section proposes an estimator of ρ and applies the jackknife method for bias reduction of $\hat{\rho}_P$ as follows:

1) Suppose we have a random sample from a bivariate normal distribution with mean vector $\mu = \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}$ and a variance covariance matrix $\Sigma = \begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}$ where $\sigma_{12} = \rho\,\sigma_1\,\sigma_2$ and it is givenby $S = ((x_1, y_1), (x_2, y_2),..., (x_n, y_n))$. In addition, an estimator of ρ is shown in Equation 2.

$$\delta(S) = \hat{\rho}_P = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

and

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i \; . \qquad (2)$$

2) The ith jackknife sample, $S_{(-i)}$, consists of the dataset with the ith observation removed. $S_{(-i)} = \left((x_1,y_1)\,,(x_2,y_2)\,,\ldots,(x_{(i-1)}\,,y_{(i-1)})\,,(x_{(i+1)},y_{(i+1)})\,,\right.$ for $i = 1, 2, \ldots, n$.

3) Let $\delta(S_{(-i)})$ be the ith jackknife replication of $\delta(S)$ and it is given by Equation 3

$$\delta\left(S_{(-i)}\right) = \hat{\rho}_{P(-i)}$$

$$= \frac{\sum_{j\neq i}^{n}(x_j - \bar{x}_{(-i)})(y_j - \bar{y}_{(-i)})}{\sqrt{\sum_{j\neq i}^{n}(x_j - \bar{x}_{(-i)})^2 \sum_{j\neq i}^{n}(y_j - \bar{y}_{(-i)})^2}} \qquad (3)$$

where

$$\bar{x}_{(-i)} = \frac{1}{n-1}\sum_{j\neq i}^{n} x_j$$

and

$$\bar{y}_{(-i)} = \frac{1}{n-1}\sum_{j\neq i}^{n} y_j$$

for $i = 1, 2, \ldots, n$.

4) Calculation of the pseudo values in the form J_i

where

$$J_i = n\delta(S) - (n-1)\delta(S_{(-i)})$$
$$= n\hat{\rho}_P - (n-1)\hat{\rho}_{P(-i)}.$$

5) The proposed estimator of ρ is given by $\hat{\rho}_J$

where

$$\hat{\rho}_J = \frac{1}{n}\sum_{i=1}^{n} J_i$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[\, n\hat{\rho}_P - (n-1)\hat{\rho}_{P(-i)} \,\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n} n\hat{\rho}_P - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\rho}_{P(-i)}$$
$$= n\hat{\rho}_P - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\rho}_{P(-i)}$$

for $\hat{\rho}_P$ and $\hat{\rho}_{P(-i)}$ are given by the format of Equations 2 and 3, respectively.

RESULTS

In order to empirically evaluate the validity and reliability of the proposed estimator, a simulation study was conducted. In the study, two populations of size 100,000, containing ordered pairs of x and y, were each generated according to a bivariate normal distribution with $\mu_1 = 2$, $\mu_2 = 4$ with an equal variance ($\sigma_1^2 = 8, \sigma_2^2 = 8$) for the first case and an unequal variance ($\sigma_1^2 = 2, \sigma_2^2 = 10$) for the second case. In addition, the correlation coefficients (ρ) of the two variables, x and y, were set at -0.9, -0.8, …, 0, 0.1, 0.2, …, 0.9, thus creating 38 populations for this simulation study. A small sample size of n=10 and large sample sizes of n=30, 50 and 60 were taken from each population by using simple random sampling with replacement with 2,000 repetitions, thus creating 152 situations for the simulation study. Then, the absolute bias and mean square error (MSE) comparisons of $\hat{\rho}_J$, $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ were performed empirically.

The simulation results presented in Figure 1 and Figure 2 reveal the absolute biases of $\hat{\rho}_J$, $\hat{\rho}_P$ and $\hat{\rho}_{OP}$. Furthermore, they confirm the bias reduction of $\hat{\rho}_P$ from the proposed estimator. For the unequal variance of two populations, the absolute bias of $\hat{\rho}_J$ was less than those of $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ for a small sample size, n=10, at all levels of the population correlation coefficient. In the case of an equal variance for the two populations, the absolute bias of $\hat{\rho}_J$ was less than those of $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ when the population correlation coefficients fell between -0.7 and 0.4. In addition, the absolute bias of $\hat{\rho}_J$ seemed to have no difference from that of $\hat{\rho}_{OP}$ when the sample sizes were not less than 30 at all levels of the population correlation coefficient for unequal and equal variances of the two populations.

Even for the large sample sizes of n=30, 50 and 60, the absolute biases of $\hat{\rho}_J$ and $\hat{\rho}_{OP}$ were less than that of $\hat{\rho}_P$ when the population correlation coefficient did not approximate zero and there was an unequal variance for the two populations, whereas where there were equal variances for the two populations, the absolute biases of $\hat{\rho}_J$ and $\hat{\rho}_{OP}$ were likely to be lower than that of $\hat{\rho}_P$ when the population correlation coefficient was positive. In addition, the absolute biases of $\hat{\rho}_J$, $\hat{\rho}_P$ and $\hat{\rho}_{OP}$

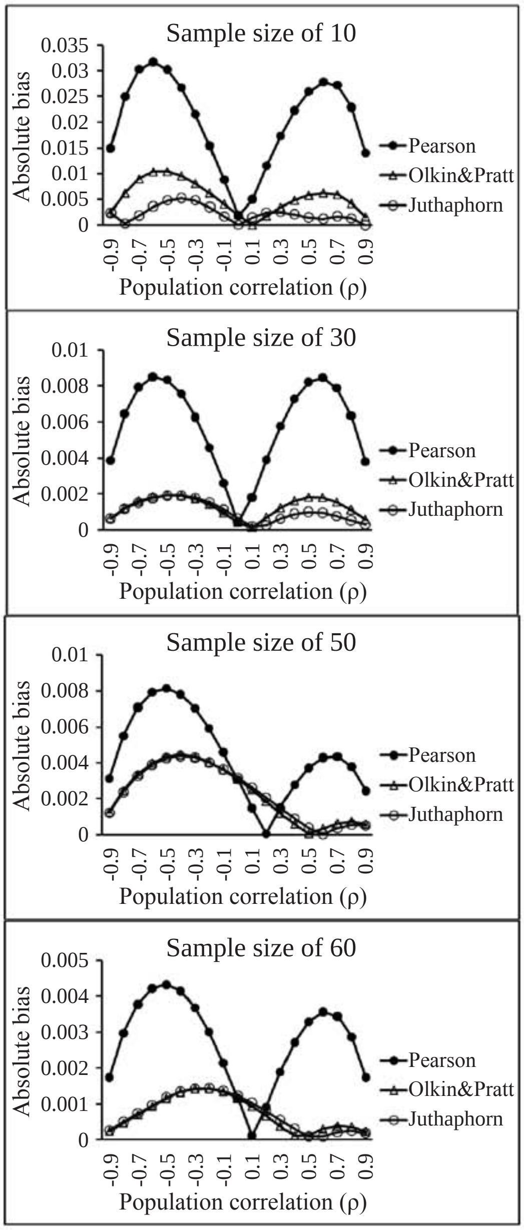

Figure 1 Absolute biases of $\hat{\rho}_J$, $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ when σ_1^2 = 2 and σ_2^2 = 10.

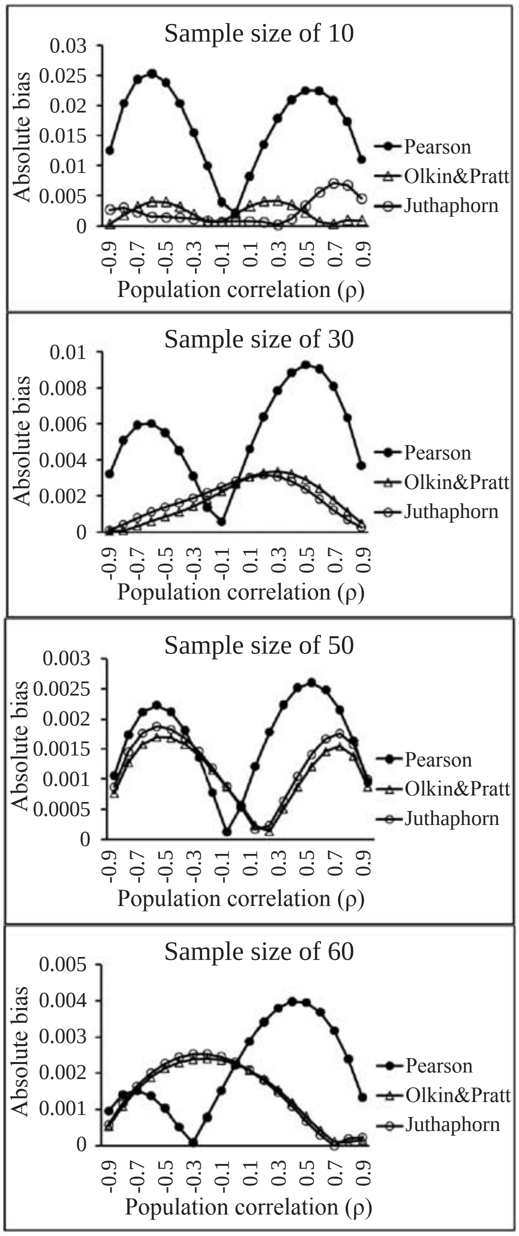

Figure 2 Absolute biases of $\hat{\rho}_J$, $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ when σ_1^2 = 8 and σ_2^2 = 8.

seemed to decrease whenever the sample size increased.

Figure 3 and Figure 4 indicate that the MSE of $\hat{\rho}_J$ seemed to have no difference from those of $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ in each situation for this study. Furthermore, the MSEs of $\hat{\rho}_J$, $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ seemed to decrease whenever the sample size increased, regardless of the population correlation coefficient.

This simulation study found that the proposed estimator, $\hat{\rho}_J$, almost completely eliminated the bias , and the performance of this estimator seemed to be better than those of $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ for the small sample size of 10.

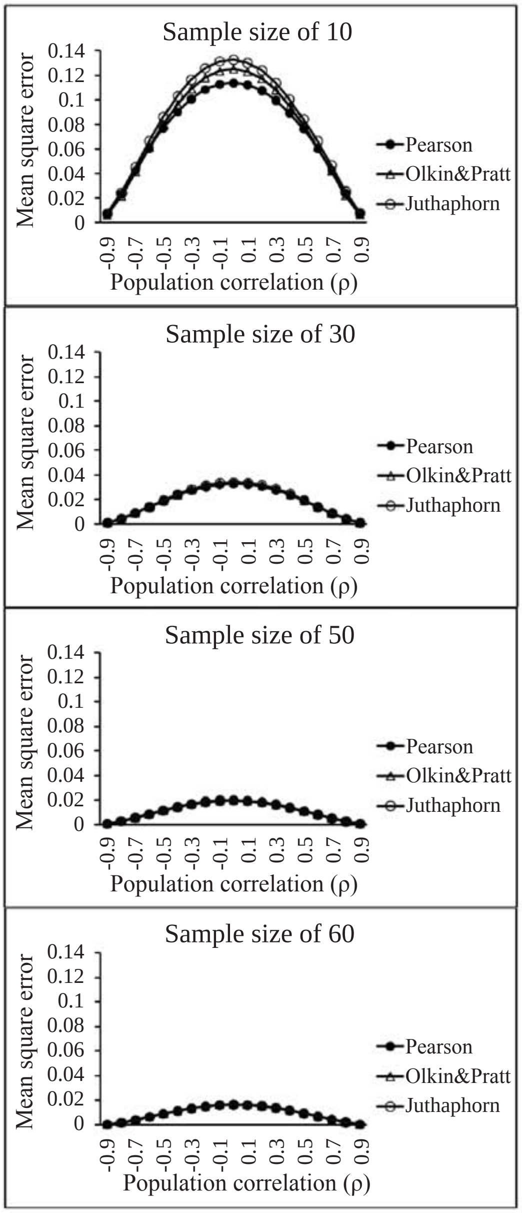

Figure 3 Mean square errors of $\hat{\rho}_J$, $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ when σ_1^2 = 2 and σ_2^2 = 10.

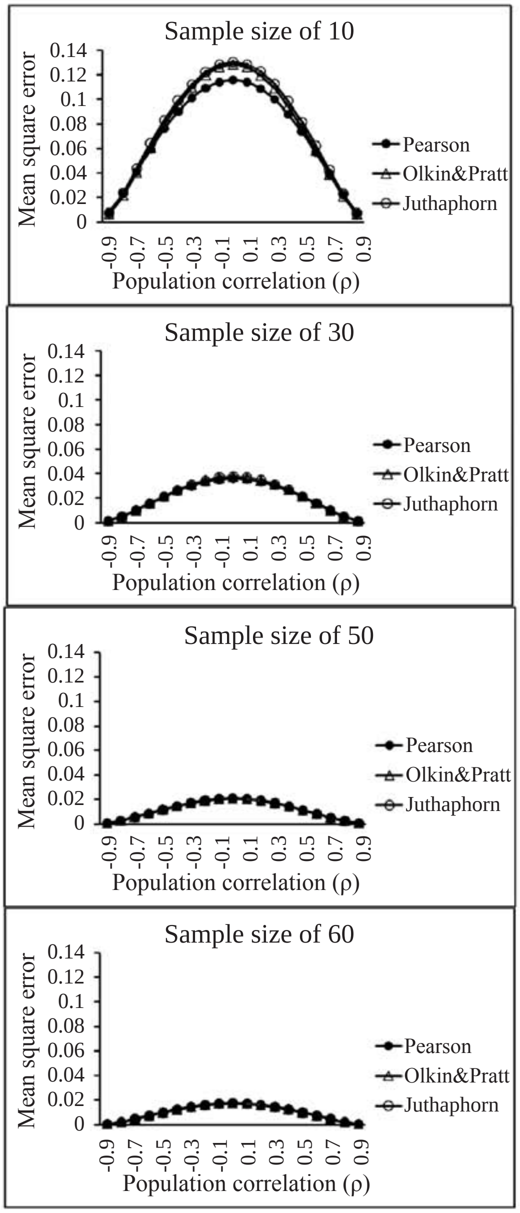

Figure 4 Mean square errors of $\hat{\rho}_J$, $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ when σ_1^2 = 8 and σ_2^2 = 8.

DISCUSSION

The simulation results showed that $\hat{\rho}_P$ seemed to be a biased estimator as mentioned by Neter *et al.* (1996) and Zimmerman *et al.* (2003). The proposed estimator, $\hat{\rho}_J$, was modified from $\hat{\rho}_P$ and the jackknife method was applied for bias reduction. The results of this simulation study showed that the bias of the proposed estimator was reduced to zero in all situations. In addition, these results also showed that the variances of two populations do not affect the bias reduction for all estimators as studied by Olkin and Pratt (1958). These findings can be applied to research in psychology, the behavioral sciences, ecology and other fields. In addition, it is possible to use computer programming to calculate $\hat{\rho}_J$ without difficulty.

CONCLUSION

This paper proposed an estimator of the population correlation coefficient for a bivariate normal distribution. The proposed estimator provided an approximately unbiased estimator of the population correlation coefficient. The results of a simulation study indicated that the performance of $\hat{\rho}_J$ seemed to be better than those $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ for a small sample size, n=10, regardless of the population correlation coefficients. In addition, the MSE of $\hat{\rho}_J$ seemed to have no difference from those of $\hat{\rho}_P$ and $\hat{\rho}_{OP}$ in each situation for this study.

ACKNOWLEDGEMENTS

The author would like to thank the Department of Statistics, Faculty of Science, Kasetsart University for the provision of necessary facilities during the research.

LITERATURE CITED

Adolph, S.C. and J.S. Hardin. 2007. Estimating phenotypic correlations: Correcting for bias due to intraindividual variability. **Funct. Ecol.** 21(1): 178–184.

Anderson, T.W. 2003. **An Introduction to Multivariate Statistical Analysis.** 3rd ed. Wiley. Hoboken, NJ, USA. 721 pp.

Efron, B. and R.J. Tibshirani. 1993. **An Introduction to the Bootstrap.** Chapman & Hall/CRC, part of the Taylor and Francis group. London, UK. 452 pp.

Fisher, R.A. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. **Biometrika** 10(4): 507–521.

__________. 1921. On the “probable error” of a coefficient of correlation deduced from a small sample. **Metron** 1: 3–32.

Gorsuch, R.L. and C.S. Lehmann. 2010. Correlation coefficients: Mean bias and confidence interval distortions. **Journal of Methods and Measurement in the Social Sciences** 1(2): 52–65.

Huson, L.W., Biostatistics Group and F.H. La-Roche. 2007. Performance of some correlation coefficients when applied to zero-clustered data. **J. Mod. Appl. Stat. Methods** 6: 530–536.

Neter, J., M.H. Kutner, C.J. Nachtsheim and W. Wasserman. 1996. **Applied Linear Statistical Models.** 4th ed.. Irwin. Chicago, IL, USA. 1,423 pp.

Olkin, I. and J.W. Pratt. 1958. Unbiased estimation of certain correlation coefficients. **Ann. Math. Statist.** 29: 201–211.

Quenouille, M.H. 1949. Approximate test of correlation in time-series. **J. R. Statist. Soc. B** 11: 68–84.

Rodgers, J.L. and W.A. Nicewander. 1988. Thirteen ways to look at the correlation coefficient. **Am. Stat.** 42(1): 59–66.

Sinsomboonthong, J. 2011a. Estimation of the correlation coefficient for a bivariate normal distribution with missing data. **Kasetsart J. (Nat. Sci.)** 45(4): 736–742.

Sinsomboonthong, J. 2011b. Jackknife maximum likelihood estimates for a bivariate normal distribution with missing data. **Journal of Thai Statistical Association** 9(2): 151–169.

Smith, C.D. and J.S. Pontius. 2006. Jackknife estimator of species richness with S-PLUS. **J. Stat. Softw.** 15: 1–12.

Tukey, J.W. 1958. Bias and confidence in not-quite large samples. **Ann. Math. Statist.** 29: 614–623.

Zimmerman, D.W., B.D. Zumbo and R.H. Williams. 2003. Bias in estimation and hypothesis testing of correlation. **Psicológica** 24: 133–158.