Kasetsart J. (Nat. Sci.) 45 : 736 - 742 (2011)

Estimation of the Correlation Coefficient for a Bivariate Normal Distribution with Missing Data

Juthaphorn Sinsomboonthong*

ABSTRACT

This study proposes an estimator of the correlation coefficient for a bivariate normal distribution with missing data, via the complete observation analysis method. Evaluation of the proposed estimator ($\hat{\rho}_J$) in comparison with the Pearson correlation coefficient ($\hat{\rho}_P$) was conducted using a simulation study. It was found that, for a higher percentage of missing data in a large sample size, the absolute bias of $\hat{\rho}_J$ was less than that of $\hat{\rho}_P$ when the population correlation coefficients (ρ) were not close to zero. In addition, the mean square error of $\hat{\rho}_J$ was not different from that of $\hat{\rho}_P$ in each situation.

Keywords: bivariate normal distribution, correlation coefficient, bias, missing data, mean square error

INTRODUCTION

Missing data, which are almost always found in research studies and caused by many possible reasons, usually introduce bias and inefficiency in parameter estimation (Little and Rubin, 2002; Norazian *et al*., 2008). Hence, well-designed data analysis is especially necessary. Principally, although incomplete data may possibly be analyzed using standard statistical methods through which missing data are ignored, an important limitation is that the methods are specifically appropriate for studies which contain small amounts of missing data. Moreover, standard statistical approaches can also cause deficiencies of data when incomplete cases are discarded. Data deficiency always causes imprecision and also an escalation in biases (Rao *et al*., 1999; Little and Rubin, 2002). Acock (2005) mentioned that this may further reduce or exaggerate statistical power. Likewise, Rotnitzky and Wypij (1994), Roth *et al*. (1996), Gorelick (2006) and Fitzmaurice (2008) mentioned that these may possibly result in invalid conclusions, since the degree of bias and the loss of precision depend not only on the fraction of complete cases and the pattern of missing data, but also on differences between the complete and incomplete cases, and the parameters of interest.

Recently, several authors have investigated problems regarding the estimation of the population correlation coefficient, ρ, for samples from a bivariate normal distribution. The maximum likelihood estimator of ρ for a bivariate normal distribution was proposed by Dahiya and Korwar (1980) for equal variances and an incomplete dataset. Garren (1998) examined the problem of maximum likelihood estimation for the correlation coefficient and its asymptotic properties in a bivariate normal model with missing data. Mudelsee (2003) studied the Pearson correlation coefficient with bootstrap confidence intervals from a bivariate climate time series and

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

Received date : 22/02/11 Accepted date : 29/06/11

unknown data distributions. In addition, the performances of Pearson correlation coefficient and Spearman correlation coefficient have been further investigated by Huson *et al*. (2007). They found that the Pearson correlation coefficient is in fact, for most practical purposes, an adequate choice for the correlation coefficient investigation. Moreover, the Pearson estimate was better than the much more widely known Spearman estimate. However, Neter *et al*. (1996) and Zimmerman *et al*. (2003) mentioned that the Pearson correlation coefficient was a biased estimator of the population correlation coefficient for bivariate normal populations. In addition, the bias decreased when the sample size increased and it was zero when the population correlation coefficients were zero and one. Furthermore, Efron and Tibshirani (1993) and Smith and Pontius (2006) have applied Jackknife's method (Quenouille, 1949, 1956; Tukey, 1958) of bias reduction to the estimation of parameters. The basic idea behind Jackknife's method lies in systematically recomputing statistics by using samples that leave out one observation at a time from the sample set. From this new set of observations, an estimator can be calculated.

Generally, in the current study, incomplete data with bivariate normal distribution were examined. The estimator of population correlation coefficient was modified from Pearson correlation coefficient, and Jackknife's method was applied for bias reduction.

MATERIALS AND METHODS

The Pearson correlation coefficient

Consider the incomplete bivariate sample from a bivariate normal distribution with mean vector (μ_1, μ_2), a variance covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ and correlation coefficient ρ.

Assume that r pairs of (X_1, X_2) are completely observed with bivariate normal distribution, but the rest n – r observations of X_2 are lost and there are only n – r observations $(0 < r < n)$ of X_1 collected (see Figure 1). All data pairs are independent and identically distributed and data are assumed to be missing completely at random (Little and Rubin, 2002). In other words, whether or not data are missing is independent of both the observed and the unobserved values of X_1 and X_2.

Based on the r data pairs, it is well-known that the maximum likelihood estimator of ρ, denoted by $\hat{\rho}_P$, is given by

$$\hat{\rho}_P = \frac{\sum_{j=1}^{r}(x_{1j}-\bar{x}_1')(x_{2j}-\bar{x}_2')}{\sqrt{\sum_{j=1}^{r}(x_{1j}-\bar{x}_1')^2 \sum_{j=1}^{r}(x_{2j}-\bar{x}_2')}} \text{ where}$$

$$\bar{x}_1' = \frac{1}{r}\sum_{j=1}^{r}x_{1j} \text{ and } \bar{x}_2' = \frac{1}{r}\sum_{j=1}^{r}x_{2j}$$

(Neter *et al*., 1996; Anderson, 2003). This estimator is often called the Pearson correlation coefficient. It is a biased estimator of ρ (unless

Observations:	**1**	**2**	**…**	**r**	**r+1**	**…**	**n**
Variable 1:	x_{11}	x_{12}	…	x_{1r}	$x_{1,r+1}$	…	x_{1n}
Variable 2:	x_{21}	x_{22}	…	x_{2r}			

Figure 1 Monotone missing data pattern for a bivariate normal distribution.

$\rho = 0$ or 1), which is usually small when sample size is large (Neter *et al*., 1996; Zimmerman *et al*., 2003).

The Jackknife's method of bias reduction

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

1) Given a sample
$X = (x_{11}, x_{12}, \ldots, x_{1r}, x_{21}, x_{22}, \ldots, x_{2r})$
and an estimator

$$\delta(X) = \hat{\rho}_P = \frac{\sum_{j=1}^{r}(x_{1j}-\bar{x}_1')(x_{2j}-\bar{x}_2')}{\sqrt{\sum_{j=1}^{r}(x_{1j}-\bar{x}_1')^2 \sum_{j=1}^{r}(x_{2j}-\bar{x}_2')}} \quad (1)$$

where $\bar{x}_1' = \frac{1}{r}\sum_{j=1}^{r}x_{1j}$ and $\bar{x}_2' = \frac{1}{r}\sum_{j=1}^{r}x_{2j}$.

2) The i^{th} Jackknife sample, $X_{(-i)}$, consists of the data set with the i^{th} observation removed.

$$X_{(-i)} = (x_{11}, x_{12}, \ldots, x_{1(i-1)}, x_{1(i+1)}, \ldots, x_{1r}, x_{21}, x_{22}, \ldots, x_{2(i-1)}, x_{2(i+1)}, \ldots, x_{2r})$$

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

3) $\delta(X_{(-i)})$ is the i^{th} Jackknife replication of $\delta(X)$ and

$$\delta(X_{(-i)}) = \hat{\rho}_{P(-i)} = \frac{\sum_{j\neq i}^{r}(x_{1j}-\bar{x}_{1(-i)}')(x_{2j}-\bar{x}_{2(-i)}')}{\sqrt{\sum_{j\neq i}^{r}(x_{1j}-\bar{x}_{1(-i)}')^2 \sum_{j\neq i}^{r}(x_{2j}-\bar{x}_{2(-i)}')^2}} \quad (2)$$

where $\bar{x}_{1(-i)}' = \frac{1}{r-1}\sum_{j\neq i}^{r}x_{1j}$ and

$\bar{x}_{2(-i)}' = \frac{1}{r-1}\sum_{j\neq i}^{r}x_{2j}$

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

4) Calculate the pseudo values in the form of J_i
where $J_i = r\delta(X) - (r-1)\delta(X_{(-i)})$

$$= r\hat{\rho}_P - (r-1)\hat{\rho}_{P(-i)}.$$

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

where $\hat{\rho}_J = \frac{1}{r}\sum_{i=1}^{r}J_i$

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

$$= \frac{1}{r}\sum_{i=1}^{r}r\hat{\rho}_P - \frac{r-1}{r}\sum_{i=1}^{r}\hat{\rho}_{P(-i)}$$

$$= r\hat{\rho}_P - \frac{r-1}{r}\sum_{i=1}^{r}\hat{\rho}_{P(-i)}$$

$\hat{\rho}_P$ and $\hat{\rho}_{P(-i)}$ are given by the format of equation (1) and (2) respectively.

RESULTS

In order to empirically evaluate the validity and reliability of the proposed estimator, a simulation study was conducted. In the study, populations of (X_1, X_2) at a size of N = 100,000 were generated in the form of a bivariate normal distribution with $\mu_1 = 2$, $\mu_2 = 3$, $\sigma_1^2 = 4$ and $\sigma_2^2 = 9$. Correlation coefficients of (X_1, X_2) at -0.9, -0.8, …, 0, 0.1, 0.2, …, 0.9 with the sample sizes of n = 20, 30 and 60 were conducted using a simple random sampling with replacement method with 2,000-times repetitions, and missing data were set at 10, 20 and 30 percentage of the total cases, thus creating 171 situations for the simulation study. Then, absolute bias and mean square error (MSE) comparisons of $\hat{\rho}_J$ and $\hat{\rho}_P$ were empirically performed.

The simulation results presented in Figure 2 reveal the absolute biases of $\hat{\rho}_J$ and $\hat{\rho}_P$. When the sample size and percentage of the missing data were 20 and 10%, respectively, the absolute bias of $\hat{\rho}_J$ was less than that of $\hat{\rho}_P$ for the population correlation coefficients (ρ) which fell between 0.1 and 0.2. For 20% missing data and the sample size of 20, the absolute bias of $\hat{\rho}_J$ was less than that of $\hat{\rho}_P$ when population correlation coefficients were not about zero.

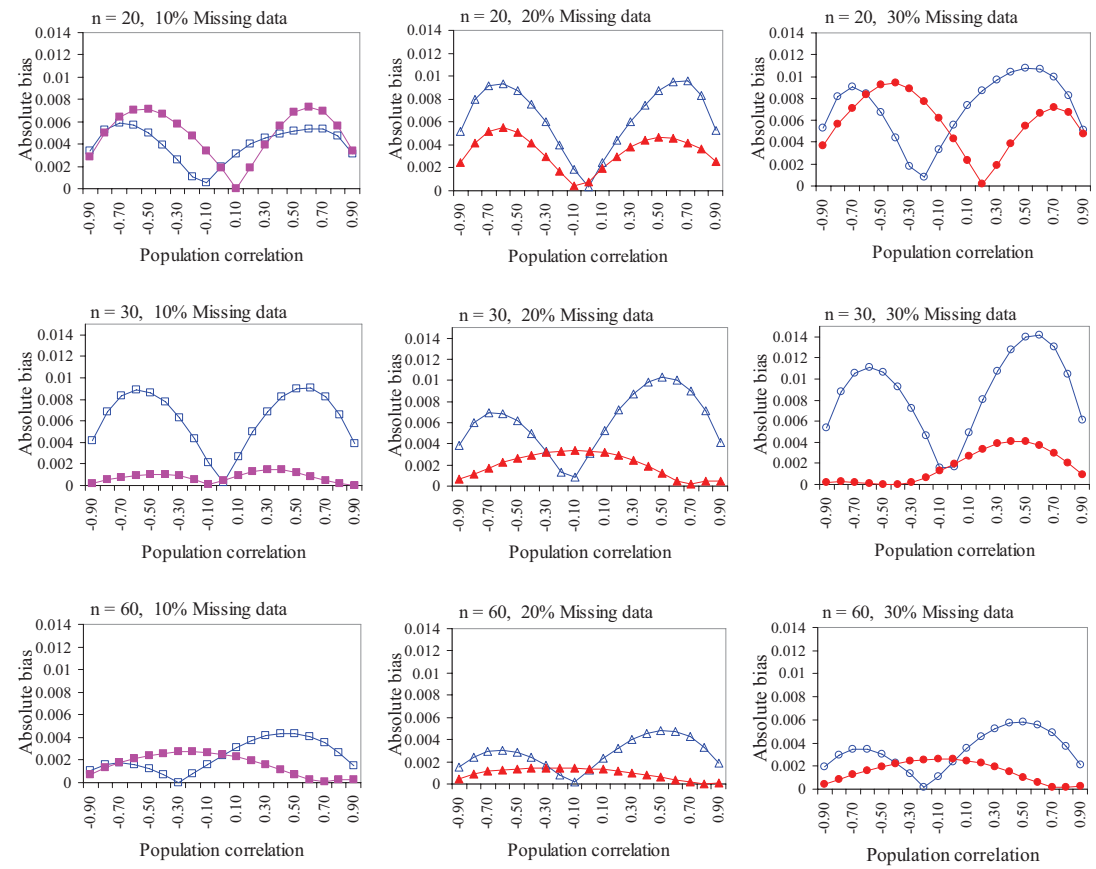

Figure 2 Comparison of absolute biases for $\hat{\rho}_J$ and $\hat{\rho}_P$ when n = 20, 30 and 60.

Likewise, the absolute bias of $\hat{\rho}_J$ was less than that of $\hat{\rho}_P$ for population correlation coefficients which were not between -0.5 and -0.1 when the percentage of missing data was 30% and the sample size was 20. In addition, the absolute bias of $\hat{\rho}_J$ was less than that of $\hat{\rho}_P$ when the sample size was greater than 20 and the percentage of missing data was greater than 10% for population correlation coefficients which were not close to zero. Moreover, the absolute biases of $\hat{\rho}_J$ and $\hat{\rho}_P$ were less than 0.004 and 0.014, respectively, for sample sizes of 30 and 60. Thus, the absolute bias of $\hat{\rho}_P$ seemed to be greater than that of $\hat{\rho}_J$ when sample sizes were 30 and 60. With a data loss of 10%, the absolute bias of $\hat{\rho}_J$ was less than that of $\hat{\rho}_P$ at all levels of population correlation coefficient for n = 30, whereas for n = 60, the absolute bias of $\hat{\rho}_J$ was less than that of $\hat{\rho}_P$ when the population correlation coefficients were positive. Moreover, the absolute biases of $\hat{\rho}_J$ and $\hat{\rho}_P$ seemed to decrease whenever the sample size increased.

Figure 3 indicates that the mean square error of $\hat{\rho}_J$ seems to have no difference from that of $\hat{\rho}_P$ in each situation for this study. Furthermore, the mean square errors of $\hat{\rho}_J$ and $\hat{\rho}_P$ seem to

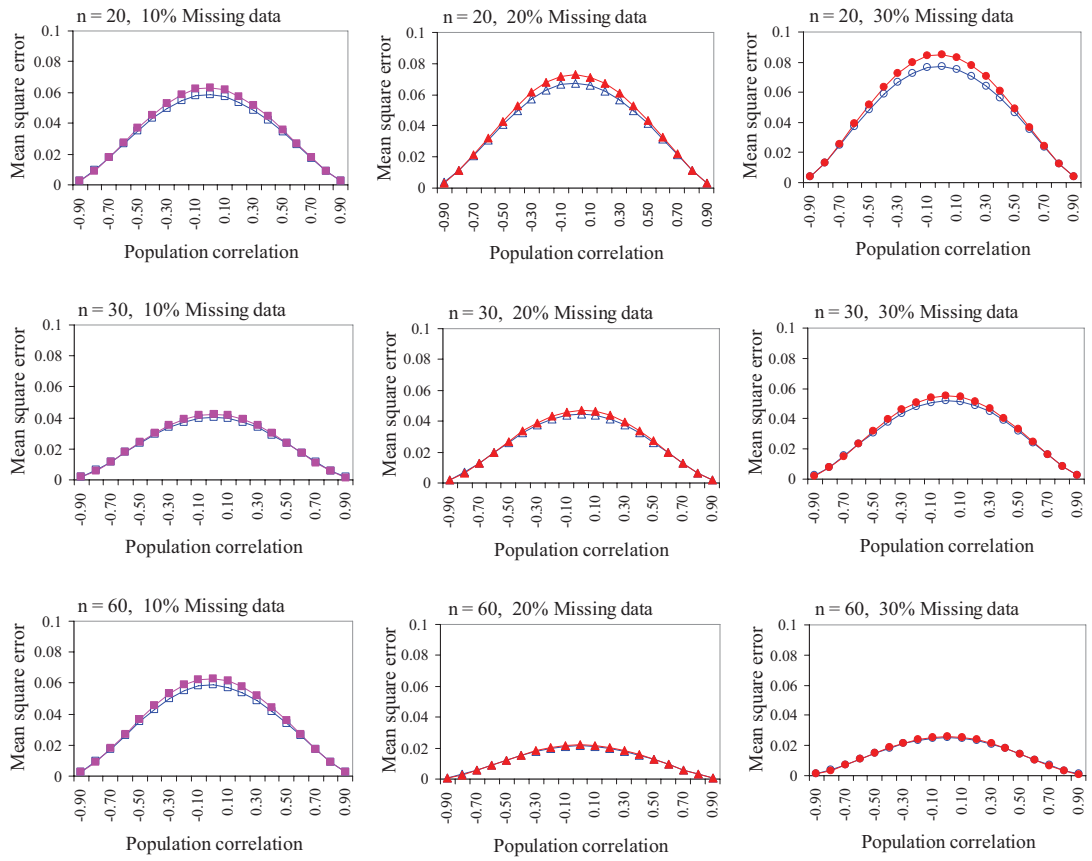

Figure 3 Mean square errors of $\hat{\rho}_J$ and $\hat{\rho}_P$ when n = 20, 30 and 60.

decrease whenever the sample size increased whatever the percentages of missing data. The ranges of the mean square errors of $\hat{\rho}_J$ and $\hat{\rho}_P$ were found to be narrower when the sample size was larger, with 20% and 30% missing data.

This simulation study found that the performance of $\hat{\rho}_J$ was better than that of $\hat{\rho}_P$ for sample sizes of 30 and 60 with a higher percentage of missing data and the population correlation coefficients were not close to zero.

DISCUSSION

The simulation results indicated that $\hat{\rho}_P$ seemed to be a biased estimator as Neter *et al.* (1996) and Zimmerman *et al.* (2003) mentioned. Hence, the bias of $\hat{\rho}_P$ can be reduced by Jackknife' s method as reported (Efron and Tibshirani, 1993; Smith and Pontius, 2006). Moreover, the bias of the proposed estimator reduced to zero for a large sample size. These findings can be applied in research, in education, psychology, medicine and other fields. Jackknife's method can be applied

in the elimination of biases in the correlation coefficient estimation for incomplete samples from bivariate normal populations. In addition, it is possible to calculate the proposed estimator without difficulty by computer programming.

CONCLUSION

This paper proposed an estimator of the correlation coefficient for a bivariate normal distribution when observations are missing from one of the variables. The proposed estimator ($\hat{\rho}_J$) was derived from the Pearson correlation coefficient ($\hat{\rho}_P$) and based on the analysis of complete cases. The results of the simulation study indicated that the absolute bias of $\hat{\rho}_J$ was less than that of $\hat{\rho}_P$ when the sample size was large for higher percentages of missing data and the population correlation coefficients were not close to zero. Furthermore, the absolute bias of $\hat{\rho}_J$ was less than 0.004 for sample sizes of 30 and 60 with whatever percentage of missing data. In addition, the mean square error of $\hat{\rho}_J$ seemed to be no different from that of $\hat{\rho}_P$ in each situation for this simulation study.

ACKNOWLEDGEMENTS

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

LITERATURE CITED

Acock, A.C. 2005. Working with missing values. **Journal of Marriage and Family** 67: 1012–1028.

Anderson, T.W. 2003. **An Introduction to Multivariate Statistical Analysis.** 3rd ed. Wiley. New Jersey. 721 pp.

Dahiya, R.C. and R.M. Korwar. 1980. Maximum likelihood estimates for a bivariate normal distribution with missing data. **The Annals of Statistics** 8: 687–692.

Efron, B. and R.J Tibshirani. 1993. **An Introduction to the Bootstrap.** Chapman& Hall/CRC. USA. 452 pp.

Fitzmaurice, G. 2008. Missing data: Implications for analysis. **Nutrition** 24: 200–202.

Garren, S.T. 1998. Maximum likelihood estimation of the correlation coefficient in a bivariate normal model with missing data. **Statistics & Probability Letters** 38: 281–288.

Gorelick, M.H. 2006. Bias arising from missing data in predictive models. **Journal of Clinical Epidemiology** 59: 1115–1123.

Huson, L.W., Biostatistics Group and F.H. La-Roche. 2007. Performance of some correlation coefficients when applied to zero-clustered data. **Journal of Modern Applied Statistical Method** 6: 530–536.

Little, R.J.A. and D.B. Rubin. 2002. **Statistical Analysis with Missing Data.** Wiley. New Jersey. 409 pp.

Mudelsee, M. 2003. Estimating Pearson's correlation coefficient with bootstrap confidence interval from serially dependent time series. **Math. Geol.** 35: 651–665.

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

Norazian, M.N., Y.A. Shukri, R.N. Azam and A.M.M. Al Bakri. 2008. Estimation of missing values in air pollution data using single imputation techniques. **ScienceAsia** 34: 341–345.

Quenouille, M.H. 1949. Approximate test of correlation in time-series. **Journal of the Royal Statistical Society. Series B (Methodological)** 11: 68–84.

Quenouille, M.H. 1956. Notes on bias in estimation. **Biometrika** 43: 353–360.

Rao, C.R., H. Toutenburg and A. Fieger. 1999. **Linear Models: Least Squares and**

Alternatives. 2nd ed. Springer-Verlag. New York. 442 pp.

Roth, P.L., J.E. Campion and S.D. Jones. 1996. The Impact of four missing data techniques on validity estimates in human resource management. **Journal of Business and Psychology** 11: 101–112.

Rotnitzky, A. and D. Wypij. 1994. A Note on the biased of estimators with missing data. **Biometrics** 50: 1 163–1170.

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

Tukey, J.W. 1958. Bias and confidence in not-quite large samples. **Annals of Mathematical Statistics** 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.