PROPERTIES OF THE SAMPLE CORRELATION OF THE BIVARIATE LOGNORMAL DISTRIBUTION

Chin-Diew Lai, Department of Statistics, Massey University, New Zealand
John C W Rayner, School of Mathematics and Applied Statistics,
University of Wollongong, , Australia
T P Hutchinson, School of Behavioural Sciences, Macquarie University, Australia

Most statistics students know that the sample correlation coefficient R is used to estimate the population correlation coefficient ρ. If the pair (X, Y) has a bivariate normal distribution, this would not cause any trouble. However, if the marginals are nonnormal, particularly if they have high skewness and kurtosis, the estimated value from a sample may be quite different from the population correlation coefficient ρ. Our simulation analysis indicates that for the bivariate lognormal, the bias in estimating ρ can be very large and it can be substantially reduced only after a large number (3-4 million) of observations. This example could serve as an exercise for the statistics students to realise some of the pitfalls in using the sample correlation coefficient to estimate ρ.

INTRODUCTION

The Pearson product-moment correlation coefficient ρ is a measure of linear dependence between a pair of random variables (X,Y). The sample (product-moment) correlation coefficient R, derived from n observations of the pair (X, Y), is normally used to estimate ρ. Historically, R has been studied and applied extensively. The distribution of R has been thoroughly reviewed in Chapter 32 of Johnson et al. (1995). While the properties of R for the bivariate normal are clearly understood, the same cannot be said about the nonnormal bivariate populations. Cook (1951), Gayen (1951) and Nakagawa and Niki (1992) obtained expressions for the first four moments of R in terms of the cumulants and cross-cumulants of the parent population. However, the size of the bias and the variance of R are still rather hazy for general bivariate nonnormal populations when $\rho \neq 0$ since the cross-cumulants are difficult to quantify in general. Although various specific nonnormal populations have been investigated, the messages on the robustness of R are conflicting.

The bivariate lognormal distribution is very well known. It arises from transforming the marginals of the bivariate normal distribution by the exponential function. It has several applications in the literature. For example, Mielke et al.(1977) use this bivariate distribution for analyses of treatment (clouding) effects on measurements (precipitation amounts) when appropriate covariates (related controlled area measurements) are available. The correlation coefficient r for the bivariate lognormal population can be obtained easily and it has been given in several books, e.g., pp. 20 of Johnson and Kotz (1972).

Results of our simulations indicate that the sample correlation *R* for the bivariate lognormal with skewed marginals and $\rho \neq 0$ has a large bias and a large variance for smaller sample sizes. One requires several million observations in order to reduce the bias and variance significantly. When we did these simulations, we got a surprise and that we still do not fully understand what is happening. Various histograms of the sample correlation coefficient *R* based on our simulation results are plotted below. Tables of summary statistics are also provided. The paper concludes with a cautionary note for students and teachers of statistics regarding the sample correlation as an estimate for ρ.

ELEMENTARY PITFALLS IN INTERPRETING THE SAMPLE CORRELATION

Let R denote the sample correlation which is an estimate of r and let r be an observed value of R. The sample correlation *R* is only a summary statistic; there are several pitfalls in interpreting it and therefore it is worthwhile emphasising these points.

- One should not confuse correlation with causation.
- $r = 0.0$ does not mean that there is no relationship between two marginals. A scatterplot might reveal a clear (though nonlinear) relationship.
- Even if the correlation is close to 1, the relationship may be obviously nonlinear.
- Many different-looking sets of points can all produce the same value of r (see Chambers et al., 1983, section 4.2, for eight scatterplots, all having $r = .7$). The well know Anscombe (1973) data has four scattered plots. All have $r = .816$.
- The value of r calculated from a small sample may be totally misleading if not viewed in the context of its likely sampling error.
- There are several other measures of statistical dependence. These include rank correlation coefficients.

Students likely have some experience in simulating sample statistics, e.g., sample mean from normal, sample mean from Cauchy and *R* from bivariate normal. They would taste a disaster in the second example. Here, we wish to alert them another potential difficulty when sampling *R* from nonnormal bivariate populations.

SAMPLE CORRELATIONS OF THE BIVARIATE LOGNORMAL

Let (X, Y) denote a pair of bivariate lognormal random variables with correlation coefficient r; derived from the bivariate normal with marginal means ζ_1 , ζ_2, standard deviations σ_1, σ_2, and correlation coefficient ρ_N

It is well known that if we start with a bivariate normal distribution, and apply any nonlinear transformations to the marginals, Pearson's product moment correlation coefficient is smaller (in absolute magnitude) in the resulting distribution than the original bivariate normal one (of course, rank correlation coefficients are unaltered provided the transformations are monotonic). The expression for the correlation coefficient of the bivariate lognormal expression can be found in page 20 of Johnson and Kotz (1972):

$$\rho = \frac{exp(\rho_N \sigma_1 \sigma_2) - 1}{\sqrt{\{exp(\sigma_1{}^2) - 1\}\{exp(\sigma_2{}^2) - 1\}}} \tag{1}$$

Eq (1) indicates that ρ is independent of ζ_1 and ζ_2, we therefore set them both to zero for convenience sake. We note that the σ's measure the skewness of the lognormal marginals : $\alpha_3 = \sqrt{\beta_1} = (\omega - 1)^{1/2}(\omega + 2), \omega = \exp(\sigma^2)$; see pp. 212 of Johnson et al. (1994).

Clearly, ρ increases as ρ_N increases. For example, for σ_1 = 1and σ_2 = 2, we have

$$\rho = \begin{cases} 0, & \text{if } \rho_N = 0 \\ 0.179, & \text{if } \rho_N = 0.5 \\ 0.666, & \text{if } \rho_N = 1 \end{cases} \tag{2}$$

We note the skewness coefficients for σ_1 = 1and σ_2 = 2 are 6.18 and 429, respectively. The correlation ρ for the bivariate lognormal may not be very meaningful if one or both of the marginals are skewed. Consider the case for which $\sigma_1 = 1$ and $\sigma_2 = 4$. By setting $\rho_N = -1$ and $\rho_N = 1$ in (2), respectively, we work out the lower and upper limits for correlation between X and Y to be -0.000251 and 0.0312. As Romano and Siegel (1986, section 4.22) say, "Such a result raises a serious question in practice about how to interpret the correlation between lognormal random variables. Clearly, small correlations may be very misleading because a correlation of 0.01372 indicates, in fact, X and Y are perfectly functionally (but nonlinearly) related."

The distribution of *R*, when (*X*, *Y*) has a bivariate normal distribution is well known and it has been well documented in Chapter 32 of Johnson and et al. (1995). The bias $(E(R) - \rho)$ and the variance of R are both of $O(n^{-1})$ and therefore ρ can be successfully estimated from samples or simulations. For nonnormal populations, the moments of *R* may be obtained from the bivariate Edgeworth expansion which involves cross-cumulant ratios of the parent population.

SIMULATION RESULTS

In order to study the sampling distribution of *R* and assess its performance as an estimator of ρ, we carried out a large-scale simulation exercise. In our simulation procedure, we use the following steps:

Step 1: Generate *n* observations from each of the pair of independent unit normals (*U*, *V*).

Step 2: Obtain the bivariate normal (X^*, Y^*) through the relationship:

$$X^* = \sigma_1 U + \zeta_1, \quad Y^* = \sigma_2 \rho_N U + \sigma_2 (1 - \rho_N^2)^{1/2} V + \zeta_2 \qquad (3)$$

Step 3: Set $X = \exp(X^*)$ and $Y = \exp(Y^*)$. Then (X, Y) has a bivariate lognormal distribution with correlation coefficient given by (2). As we are only interested in the correlation coefficient, we set both ζ_1 and ζ_2 to zero.

All the simulations and plottings are carried out using *MINITAB* commands: (i) Simulate $U \to C1,\ V \to C2$, (ii)$C1 * \sigma_1 \to C3,\ \sigma_2 * \rho_N * C1 + \sigma_2 * \sqrt{1 - \rho_N^2} * C2 \to C4$ (iii) $\exp(C3) \to C5, \exp(C4) \to C6$ and (iv) $Corr\ \mathrm{C5\text{-}\ C6} \to \mathrm{M1}$ (Here, C stands for ‘column’ whereas M stands for ‘matrix’).

Three cases are considered: (i) $\rho_N = 1$, (ii) $\rho_N = 0.5$ and (iii) $\rho_N = 0$, each with $\sigma_1 = 1$and $\sigma_2 = 2$; and their corresponding correlation coefficients of the bivariate lognormal population are (i)$\rho = 0.666$, (ii)$\rho = 0.179$ and (iii)$\rho = 0$, respectively. The following histograms are plotted for the three cases considered:

Fig 1: 50 Samples of 4 Million (with rho = 1)

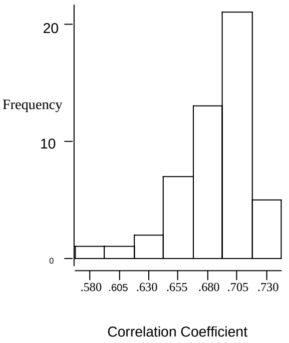


Fig 2: 100 Samples of 3 million (rho =0.5)

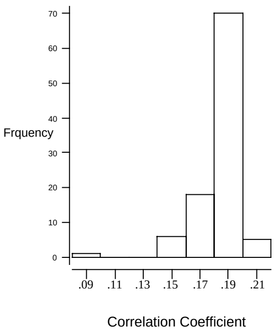


Fig 3: 100 Samples of 3 Million (rho = 0)

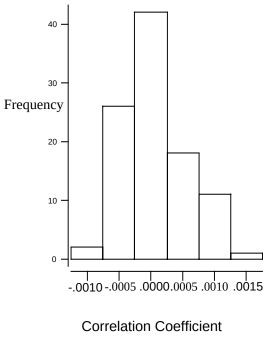


Table 1. Summary of Simulations

	ρ	Sample Size	No of Samples	Mean	Standard Deviation
Fig 1	0.666	4 Million	50	0.68578	0.029979

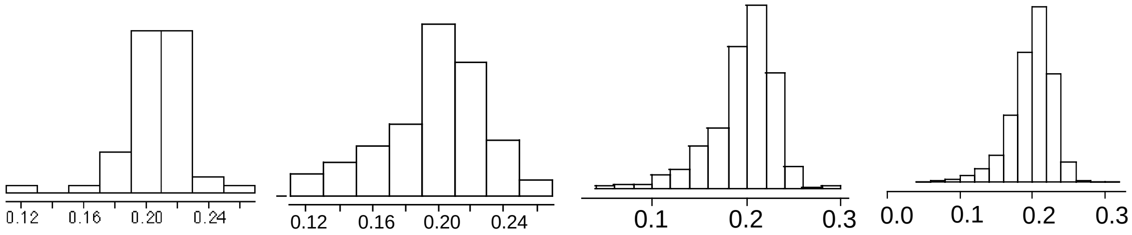
Fig 2	0.179	3 Million	100	0.18349	0.015889
Fig 3	0	3 Million	100	0.00006	0.000526

The plots displayed above indicate that the distributions of *R* are skewed to the left, and, except for the case ρ = 0, they have quite large variances even for such large sample sizes. We have also calculated the asymptotic expansions for both the bias and the variance of *R*, and found, except when ρ = 0, the leading coefficients in each case to be very large. So there is sound theory behind the simulation demonstrations.

For the bivariate normal, the bias in *R* as an estimate of ρ is approximately $-\rho(1-\rho^2)n^{-1}/2$ and that $\mathrm{var}(R)\approx\left(1-\rho^2\right)^2 n^{-1}$ (pp. 556, Johnson et al., 1995). So for these values of r (i.e., r = 0.666, 0.179 and 0), we would expect the standard errors to be 0.0003, 0.0006, and 0.0005, respectively.

In order to reassure the readers that 50 or 100 samples is sufficient, we now let the number of simulations, k say, varies, but fix the sample size to n=100,000 of case (ii). The following histograms indicate that the shape changes very little as k varies; all are skewed to the left.

Figure 4. Histograms of *R* with ρ_N = 0.5, n =100,000 (ρ = 0.179, σ_1=1,σ_2=2)



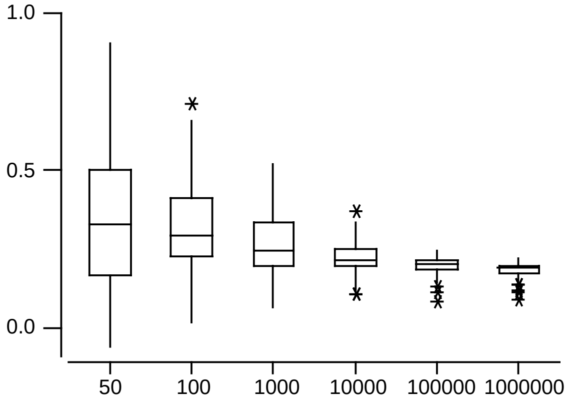
(a) k = 50 (b) k = 100 (b) k = 500 (c) k = 1000

Table 2. Summary Statistics from 4 values of k (n = 100,000)

k	Mean	Median	StDev	Min	Max	Q1	Q3
50	0.20644	0.20875	0.01978	0.12860	0.25300	0.19606	0.21871
100	0.19779	0.20372	0.03118	0.11145	0.25910	0.18012	0.22023
500	0.19582	0.20131	0.03460	0.04924	0.28928	0.18159	0.21855
1000	0.19931	0.20418	0.02965	0.04472	0.30495	0.18370	0.21913

On the other hand, if we fix k = 100 but allow n to varies from n = 50 to n = 1000000, we then have the following box-plots together with a table of summary statistics.

Figure 5. Box-plots of various n, k=100   Table 3. Summary (ρ = 0.179, σ_1=1,σ_2=2)

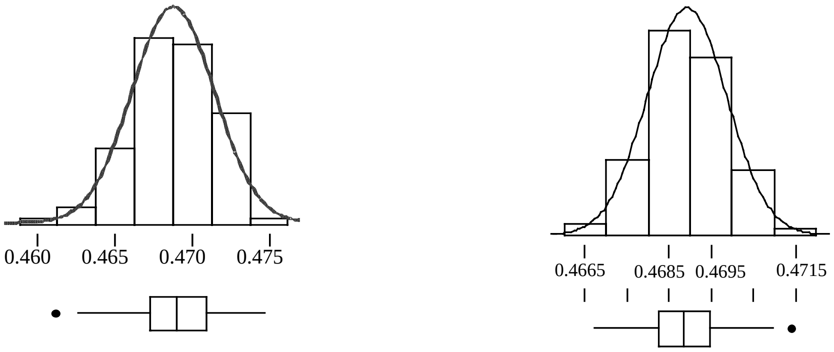


Sample Size	Mean	Median	StDev
50	0.3379	0.3289	0.2237
100	0.3258	0.2928	0.1506
1000	0.2613	0.2421	0.0960
10000	0.2195	0.2140	0.0501
100000	0.1979	0.2015	0.0266
1000000	0.1849	0.1905	0.0214

The last column of the preceding table indicates that the stand error is not proportional to $n^{-1/2}$ as one would probably expect should this be a well-behaved bivariate distribution..

If the skewness of the marginals is reduced, *R* seems to become more robust. For example, consider the case when $\sigma_1 = \sigma_2$ =.5 and $\rho_N = 0.5$; so that ρ =.4688 (recall, σ's measure the skewness of the lognormals). 100 samples of (a) 100,000 and (b) 1million observations were simulated, and their results are now summarized as follows:

Figure 6. Histograms based on 100 Samples (with $\sigma_1 = \sigma_2$ =.5 and ρ =.4688)



(a) Sample size : 100,000

(b) Sample size:1 million

Table 4: Summary Statistics of Two Different Sample Sizes (k =100)

	Mean	StDev	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Fig 4a	0.4689	0.0027	0.4611	0.4672	0.4689	0.4710	0.4747
Fig 4b	0.4689	0.0009	0.4667	0.4682	0.4688	0.4695	0.4714

We see that the normals fit the above data well. Indeed, formal goodness of fit tests show almost perfect fit. It seems that the skewness of the marginals affect the skewness of *R*.

CONCLUSION

Many non-normal bivariate distributions are of concern in engineering, geology, and meteorology, as well as in psychology. Often, it is necessary to estimate the correlation coefficient from the sample correlation *R*. In most cases the sample sizes concerned are in the order of hundreds instead of thousands or millions for obvious reasons. So the bias may be quite significant in some cases, especially if ρ is not close to zero.

By using an easily understood example we have illustrated the problem in estimating the population correlation and thereby we lend support to the claim of the non-robustness of *R*. To our knowledge, most elementary texts do not discuss or highlight this important issue. It is our view that undergraduates in statistics should be adequately cautioned about this problem and be encouraged to check for the underlying assumptions on the populations before reporting their findings on the correlation.

REFERENCE

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician, 27,* 17-20

Chambers, J. M., Cleveland, W S, Kleenex, B and Tukey, P A (1983). *Graphical Methods for Data Analysis*.. Belmont, California: Wadsworth, and Boston: Duxbury.

Cook, M. B. (1951) Bi-variate k-Statistics and Cumulants of Their Joint Sampling Distribution. *Biometrika, 38,* 179-195

Gayen, A. K. (1951) The Frequency Distribution of the Product-Moment Correlation Coefficient in Random Samples of Any Size from Non-Normal Universe. *Biometrika, 38*, 219-247.

Johnson, N. L. and Kotz, S. and Balakrishnan, N. (1994). *Distributions in Statistics: Continuous Univariate Distributions* , Vol 1, Second Edition. New York, Wiley.

Johnson, N. L. and Kotz, S. and Balakrishnan, N. (1995). *Distributions in Statistics: Continuous Univariate Distributions* , Vol 2, Second Edition. New York, Wiley.

Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*.. New York, Wiley.

Mielke, P. W., Williams, J. S. and Wu, S. C. (1977). Covariance Analysis Techniques Based on Bivariate Log-Normal Distribution with Weather Modification Applications. *Journal of Applied Meteorology, 16***,** 183-187.

Nakagawa, S. and Niki, N. (1992). Distribution of Sample Correlation Coefficient for Nonnormal Populations. *Journal of Japanese Society of Computational Statistics, 5*, 1-19.

Romano, J. P. and Siegel A. F. (1986). Counter Examples in Probability and Statistics.. Monterey, California: Wadsworth and Brooks/Cole.