© Journal of Applied Mathematics & Decision Sciences, 3(1), 7- 19 (1999)
Reprints available directly from the Editor. Printed in New Zealand

ROBUSTNESS OF THE SAMPLE CORRELATION - THE BIVARIATE LOGNORMAL CASE

C D LAI
Statistics, IIST, Massey University, New Zealand C.Lai@massey.ac.nz

J C W RAYNER
School of Mathematics and Applied Statistics, University of Wollongong, Australia

T P HUTCHINSON
School of Behavioural Sciences, Macquarie University, Australia.

Abstract. The sample correlation coefficient R is almost universally 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.

The bivariate lognormal is chosen as our case study for this robustness study. Two approaches are used: (i) by simulation and (ii) numerical computations.

Our simulation analysis indicates that for the bivariate lognormal, the bias in estimating ρ can be very large if $\rho \neq 0$, and it can be substantially reduced only after a large number (three to four million) of observations. This phenomenon, though unexpected at first, was found to be consistent to our findings by our numerical analysis.

Keywords: Asymptotic Expansion, Bias, Bivariate Lognormal, Correlation Coefficients, Cumulant Ratio, Nonnormal, Robustnes, Sample Correlation.

1. 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 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 Johnson et al. (1995, pp.580) remarked that "Contradictory, confusing, and uncoordinated floods of information on the 'robustness' properties of the sample correlation coefficient R are scattered in dozens of journals." We do not intend to enter into the fray. Instead we use an easily understood example to illustrate that we do have a problem in estimating the population correlation when $\rho \neq 0$ and when the skewness of the marginal populations is large.

Results of our simulations indicate that for smaller sample sizes the sample correlation R for the bivariate lognormal with skewed marginals and with $\rho \neq 0$ has large bias and large variance. Several million observations are required in order to reduce the bias and variance significantly. This result may be surprising to many readers. Various histograms of the sample correlation coefficient R based on our simulation results are plotted Section 3. Tables of summary statistics are also provided.

The present study is in some way a follow-up of Hutchinson (1997) who noted that the sample correlation is possibly a poor estimator of ρ. The advantage of using the bivariate lognormal as our case study on robustness of R, apart from the ease of simulations, is the tractability of the cross-cumulants. We compute the lower cross-cumulant ratios that contribute to the terms in n^{-1} and n^{-2} in the mean and variance of R. The magnitude of these values cause difficulties in estimating the sizes of the bias and variance. Nevertheless, they do shed some light on where the difficulties lie.

2. Sample Correlation of the Bivariate Lognormal Distribution

Let (X, Y) denote a pair of bivariate lognormal random variables with correlation coefficient ρ, 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 in the resulting distribution is smaller in absolute magnitude than the original bivariate normal one. Of course, if the transformations are monotonic, rank correlation coefficients are unaltered. The expression for the correlation coefficient of the bivariate lognormal can be found in Johnson and Kotz (1972, p.20):

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

Since (1) indicates that ρ is independent of ζ_1 and ζ_2, we subsequently 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 Johnson et al. (1994, pp. 212). Eq(1) indicates that ρ increases as ρ_N increases for any fixed σ_1 and σ_2.

For $\sigma_1 = 1$ and $\sigma_2 = 2$, we have

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

We note the skewness coefficients for $\sigma_1 = 1$and $\sigma_2 = 2$ 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 (1), 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 has been well documented in Johnson and et al. (1995, Chapter 32). 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.

3. Simulation Study

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:

$$X^* = \sigma_1 U_1, \quad Y^* = \sigma_2 \rho_N U + \sigma_2 (1-\rho_N^2)^{1/2} V \qquad (3)$$

Step 3: Set $X = \exp(X^*)$ and $Y = \exp(Y^*)$. Then (X, Y) has a bivariate lognormal distribution with correlation coefficient given by (1). As we are only interested in the correlation coefficient, we set both ζ_1 and ζ_2 to zero.

All the simulations and plots are carried out using *MINITAB* commands.

Three cases are considered, each with $\sigma_1 = 1$ and $\sigma_2 = 2$: (i) $\rho_N = 1$, (ii) $\rho_N = 0.5$ and (iii) $\rho_N = 0$. The 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)

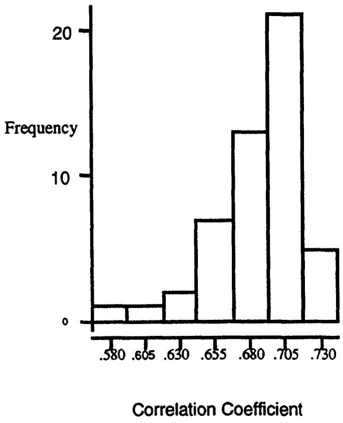


Fig 2: 100 Samples of 3 million (rho =0.5)

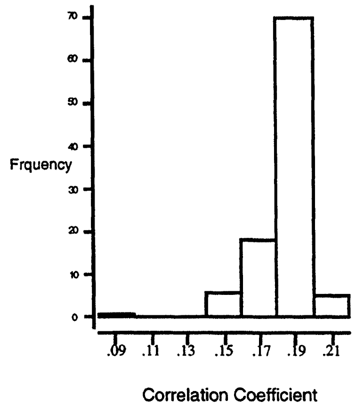


Fig 3: 100 Samples of 3 Million (rho = 0)

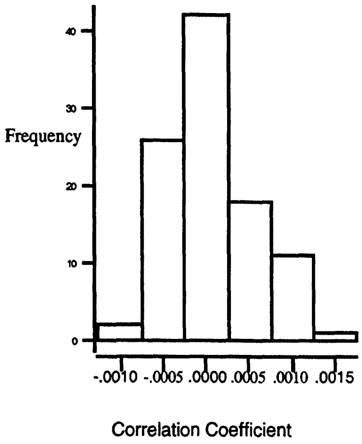


Table 1: *Summary of Simulations*

	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	0.0	3 Million	100	0.00006	0.000526

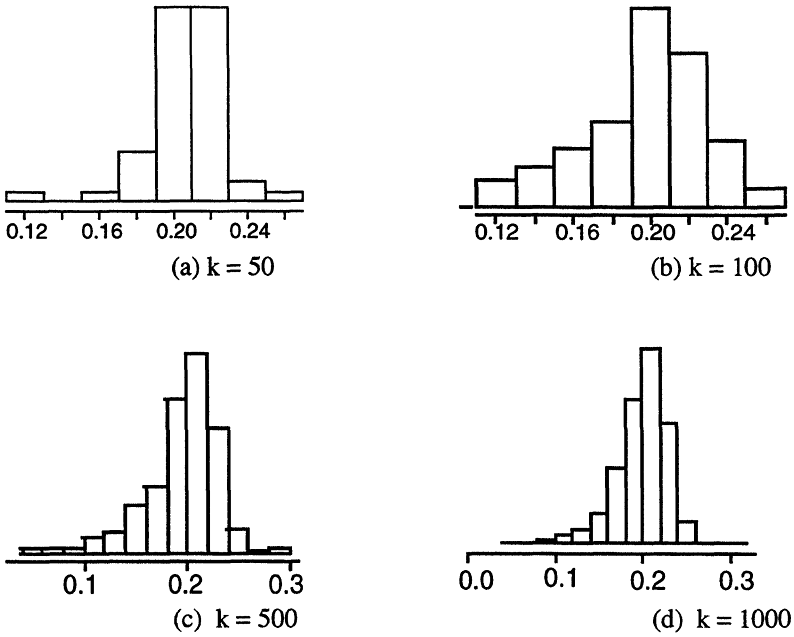
The plots displayed above indicate that the distributions of R are skewed to the left, and, except for the case $\rho = 0$, they have quite large variances even for such large sample sizes. We have also calculated the asymptotic expansions for both the bias and the variance of R, and found, except when $\rho = 0$, 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 $\mathrm{var}(R) \approx (1-\rho^2)^2/n$; see Johnson et al. (1995, pp. 556). So for ρ = 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 consider case (ii) with fixed the sample size n = 100,000 and let the number of simulations, k say, vary.

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 $\rho_N = 0.5$, ($\rho = 0.179$), $\sigma_1 = 1$, $\sigma_2 = 2$, n = 100,000)



(a) k = 50

(b) k = 100

(c) k = 500

(d) k = 1000

Table 2: Summary Statistics from four values of k, all with n = 100,000

k	Mean	Median	St Dev	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 and allow n to vary from n = 50 to n = 1000000, we then have the following box-plots and table of summary statistics:

Figure 5: Box-plots of various n, k = 100

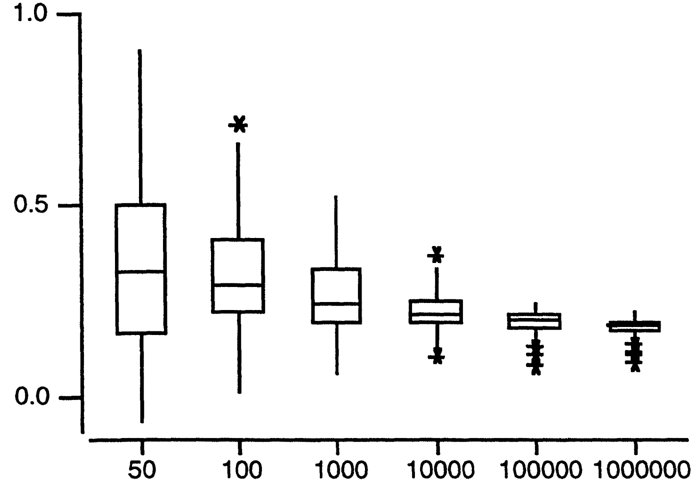


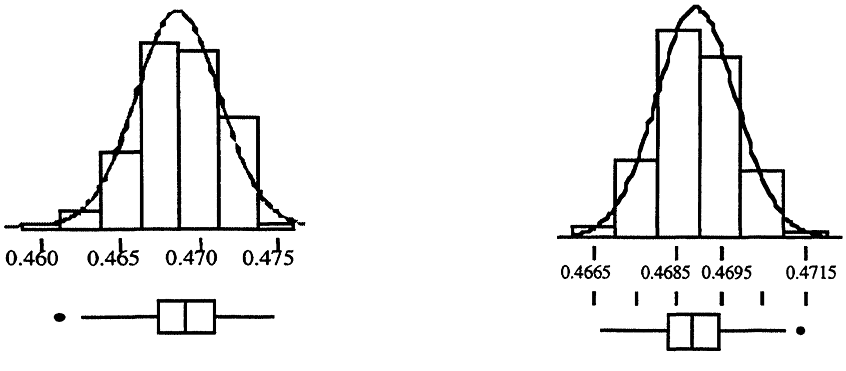
Table 3: Summary:(ρ =0.179, σ_1=1, σ_2=2)

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 suggests that the standard 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, the bias and sampling variance of R both seem to be reduced. For example, consider the case when $\sigma_1 = \sigma_2 = 0.5$ and $\rho_N = 0.5$; then $\rho = 0.4688$. Also recall that the σ's measure the skewness of the lognormals. We simulated 100 samples of (a) 100,000 and (b) 1million observations, and their results are now summarized as follows:

Fig 6: Histograms based on 100 Samples (with $\sigma_1 = \sigma_2 = 0.5$ and $\rho = 0.4688$)



(a) Sample size: 100,000

(b) Sample size: 1 million

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

	Mean	StDev	Min	1st Quartile	Median	3rd Quartile	Max
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.

4. Cross-Cumulant Ratios and Asymptotic Expansions

4.1. Asymptotic Expansions

The expected value of R for any bivariate distributions (not necessarily bivariate normal) with the population correlation coefficient ρ, is given by (see Johnson and et al 1995, page 562):

$$E(R) = \rho + \frac{1}{n}\left[-\rho(1-\rho^2)/2 + \frac{1}{8}L_{4,1}\right] + O(n^{-2}) \tag{4}$$

where

$$L_{4,1} = 3\rho(\gamma_{40} + \gamma_{04}) - 4(\gamma_{31} + \gamma_{13}) + 2\rho\gamma_{22}, \tag{5}$$

and γ_{ij} denotes the cumulant ratio

$$\gamma_{ij} = \kappa_{ij}\kappa_{20}^{-i/2}\kappa_{02}^{-j/2} \tag{6}$$

with κ_{ij}being the cross cumulants of the bivariate population.

Nakagawa and Niki (1992) include an extra term $4\gamma_{11}^3$ in (5), but we accept the expression given by Johnson et al (1995). Nevertheless, we subsequently find the additional term is numerically small and it makes no difference to our calculations.

Equation (4) may be written alternatively as:

$$E(R)-\rho=\frac{1}{n}\left[-\rho(1-\rho^2)/2+\frac{1}{8}L_{4,1}\right]+O(n^{-2}) \qquad (7)$$

Rodriguez (1982) noted that R is an approximately unbiased as well as a consistent estimator of ρ.

Gayen (1951) showed that after adding the term in n^{-2}, (7) then becomes:

$$\begin{aligned}E(R)-\rho &= \frac{1}{n}\left[-\rho(1-\rho^2)/2+\frac{1}{8}L_{4,1}\right]\\ &+\frac{1}{n^2}\left\{\frac{3}{8}\rho(1-\rho^2)(1+3\rho^2)+\frac{15}{16}\rho(1-\rho^2)(\gamma_{40}+\gamma_{04})+\frac{1}{4}(9\rho^2-5)(\gamma_{31}+\gamma_{13})\right.\\ &\left.-\frac{1}{8}\rho(9\rho^2+7)\gamma_{22}+\frac{5}{4}\rho(\gamma_{30}^2+\gamma_{03}^2)+\frac{3}{4}\rho(\gamma_{21}^2+\gamma_{12}^2)-\frac{3}{2}(\gamma_{30}\gamma_{21}+\gamma_{03}\gamma_{12})-\gamma_{21}\gamma_{12}\right\}\\ &+O(n^{-3}) \qquad (8)\end{aligned}$$

The variance of R is given by:

$$Var\{R\}=\frac{1}{n}\left\{\left(1-\rho^2\right)^2+\frac{1}{4}L_{4,2}\right\}+O(n^{-2}) \qquad (9)$$

where

$$L_{4,2}=\rho^2\left(\gamma_{40}+\gamma_{04}\right)-4\rho\left(\gamma_{13}+\gamma_{31}\right)+2(2+\rho^2)\gamma_{22} \qquad (10)$$

4.2. Cumulants and Cross Cumulant Ratios

The bivariate cumulants may be expressed in terms of product moments μ'_{ij} (about the origin) (See Cook (1951)):

$$\begin{aligned}\kappa_{30}&=\mu'_{30}-3\mu'_{20}\mu'_{10}+\mu'^3_{10}\\ \kappa_{21}&=\mu'_{21}-\mu'_{20}\mu'_{01}-2\mu'_{11}\mu'_{10}+2\mu'^2_{10}\mu'_{01}\\ \kappa_{40}&=\mu'_{40}-4\mu'_{30}\mu'_{10}-3\mu'^2_{20}+12\mu'_{20}\mu'^2_{10}-6\mu'^4_{10}\\ \kappa_{31}&=\mu'_{31}-\mu'_{30}\mu'_{01}-3\mu'_{21}\mu'_{10}-3\mu'_{20}\mu'_{11}+6\mu'_{20}\mu'_{10}\mu'_{01}+6\mu'_{11}\mu'^2_{10}-6\mu'^3_{10}\mu'_{01}\end{aligned}$$

$$\kappa_{22} = \mu'_{22} - 2\mu'_{21}\mu'_{01} + 2\mu'_{20}\mu'^{2}_{01} - \mu'_{20}\mu'_{02} - 2\mu'_{12}\mu'_{10} - 2\mu'^{2}_{11}$$
$$+ 8\mu'_{11}\mu'_{10}\mu'_{01} - 6\mu'^{2}_{10}\mu'^{2}_{01} + 2\mu'^{2}_{10}\mu'_{02}$$

When $\xi_1 = \xi_2 = 0$, using Johnson and Kotz (1972, page 20), the product moments of the bivariate lognormal distribution are

$$\mu'_{ij} = \exp\left\{ ij\rho_N\sigma_1\sigma_2 + \frac{1}{2}\left\{ i^2\sigma_1^2 + j^2\sigma_2^2 \right\} \right\} \tag{11}$$

In what follows, we assume and $\sigma_1 = 1$and $\sigma_2 = 2$ so that

$$\mu'_{ij} = \exp(ij\rho_N)\exp\left\{ \frac{1}{2}\left\{ i^2 + 4j^2 \right\} \right\} \tag{12}$$

Using (12), the cross-cumulants formulae, and (6) we obtained Table 5 and Table 6 below:

Table 5: Cross-Cumulants Ratios

ρ	γ_{03}	γ_{30}	γ_{21}	γ_{12}	γ_{22}	γ_{40}	γ_{04}
0	414.36	8.47	0	0	0	110.94	9220556
.179	414.36	8.47	0.36	4.78	78.43	110.94	9220556
.666	414.36	8.47	1.11	4.91	4734.83	110.94	9220556

Table 6: Cross-Cumulant Ratios and *L* functions

ρ	γ_{13}	γ_{31}	$L_{4,1}$	$L_{4,2}$
0	0	0	0	0
.179	6036.47	19.86	4927300.88	291421.80
.666	127316.53	462.84	18956928	3772568.34

Now

$$\gamma_{11} = \begin{cases} 0, & \text{for } \rho_N = 0 \\ 0.07, & \text{for } \rho_N = .5 \\ 0.18, & \text{for } \rho_N = 1 \end{cases}$$

After adding $4\gamma_{11}^3$, $L_{4,1}$ now becomes

$$L_{4,1} = \begin{cases} 0, & \text{for } \rho = 0 \\ 4927300.88, & \text{for } \rho = .179 \\ 18956928.02, & \text{for } \rho = .666 \end{cases}$$

which shows that the additional term hardly makes any change to our calculations.

Table 7 below gives the values of the appropriate terms in (8) and (9), which we obtained from the last two tables:

Table 7: Coefficients in the Asymptotic Expansions for the Mean and Variance

ρ	coeff of n^{-1} [bias (8)]	coeff of n^{-2} [bias (8)]	coeff of n^{-1} [var (9)]
0	1	0	1
.179	615912.44	3311342.19	72856.39
.666	2369615.63	3303158.59	943142.40

We note that the values of the last two rows of Table 7 very large. It is our conjecture that further coefficients in the asymptotic expansion of the bias and variance are large also. This is why the expansion up to the order n^{-2} cannot be used to estimate the bias and variance when $\rho \neq 0$.

Conclusion

Many non-normal bivariate distributions are of interest in engineering, geology, meteorology, and psychology. Often the correlation coefficient is used to estimate 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 that in estimating the population correlation by the sample correlation, a large bias and a large sampling variance may occur. We therefore lend our support to the claim that R is not a robust estimator of the population correlation. To our knowledge, most elementary texts do not discuss or highlight this important issue. It is our view that statistics students, as well as researchers, should be cautioned about this problem. The underlying assumptions on the populations should be checked before reporting their findings on the correlation. _

Acknowledgment

This paper is an expanded version of Lai et al. (1998) which appeared in Volume 1 of the Proceedings of ICOTS 5. The authors are grateful to the ISI for their permission to produce this paper in the current form.

References

1. Cook, M. B. (1951). Bi-variate k-Statistics and Cumulants of Their Joint Sampling Distribution. *Biometrika*, Vol 38, pp.179-195
2. Gayen, A. K. (1951). The Frequency Distribution of the Product-Moment Correlation Coefficient in Random Samples of Any Size from Non-Normal Universe. *Biometrika*, Vol 38, pp. 219-247.
3. Hutchinson, T. P. (1997). A Comment on Correlation in Skewed Distributions, *The Journal of General Psychology*, Vol 124(2), pp 211-215
4. Johnson, N. L. and Kotz, S. and Balakrishnan, N. (1994). *Distributions in Statistics: Continuous Univariate Distributions* , Vol 1, Second Edition. New York, Wiley.
5. Johnson, N. L. and Kotz, S. and Balakrishnan, N. (1995). *Distributions in Statistics: Continuous Univariate Distributions* , Vol 2, Second Edition. New York, Wiley.
6. Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions..* New York, Wiley.
7. Lai, C. D. , Rayner, J. C. W., and Hutchinson, T. P. (1998). Properties of the Sample Correlation Coefficient of the bivariate Lognormal distributions. *Proceedings of the Fifth International Conference on Teaching of Statistics* (ICOTS 5), 21-26 June 1998, Singapore. Editors: L. Pereira-Mendoza etc, Vol 1, pp 309-315, International Statistical Institute.
8. Nakagawa, S. and Niki, N. (1992) Distribution of Sample Correlation Coefficient for Nonnormal Populations. *Journal of Japanese Society of Computational Statistics*, Vol 5, pp.1-19.
9. Romano, J. P. and Siegel A. F. (1986). Counter Examples in Probability and Statistics.. Monterey, California: Wadsworth and Brooks /Cole.
10. Rodriguez, R. N. (1982). Correlation. In *Encyclopedia of Statistical Sciences.* Vol **3**, 194-204, New York, Wiley.