Equation Chapter 1 Section 1

SPURIOUS CORRELATIONS FOR STATIONARY AR(1) PROCESSES

Christos Agiakloglou* and Apostolos Tsimpanos**

*Department of Economics, University of Piraeus, 80 Karaoli & Dimitriou Street, 18534 Piraeus, Greece

**Department of Statistics & Insurance Science, University of Piraeus, 80 Karaoli & Dimitriou Street, 18534 Piraeus, Greece

Abstract

Spurious correlations occur when two independent time series are found to be correlated according to the typical statistical procedure of testing the null hypothesis of zero correlation in the population. This study, using a Monte Carlo analysis, examines the spurious correlation phenomenon for two independent stationary AR(1) processes and it finds that the source of this behavior is in the distribution of the relevant t statistic. However, using the variance of the sample correlation coefficient of two independent stationary AR(1) processes, suggested by Bartlett (1935), this study finds no evidence of spurious correlations for moderate and large sample sizes.

Keywords: Correlation coefficient, spurious correlations, stationary AR(1) processes and non-stationary processes

JEL Classification: C22

*Corresponding author: Tel.: +30-210-414-2290, Fax.: +30-210-414-2290
E-mail address: agiaklis@unipi.gr

1. Introduction

In many economic applications interest is focused on investigating the existence of a linear relationship between two random variables using the sample correlation coefficient which declares the strength of the linear association in the sample and can be used as the basis of testing the null hypothesis of no linear correlation in the population. In some cases, however, an analyst may get misleading statistical results with respect to the degree and to the existence of a linear association between two random variables, when indeed these two random variables are not linearly related. This phenomenon is known as the spurious correlation phenomenon, introduced by Yule (1926), although in the literature it received a considerable attention as the spurious regression phenomenon presented by the pioneer work of Granger and Newbold (1974), in which, using a Monte Carlo analysis, they showed that the regression of two independent non-stationary time series can generate spurious results. Furthermore, this phenomenon was well-documented mathematically by Phillips (1986) and extended by Granger *et al* (2001) for two stationary independent AR(1) processes.

Spurious correlations and spurious regressions are two similar if not identical terms referring to the same phenomenon of obtaining false evidence about the existence of a linear relationship between two variables. However, there is major difference between these two terms with respect to the ability of detecting such behavior. In the case of spurious correlations the analyst has no indication about the existence of such behavior, unless he or she has some *a priori* information about the relationship between these two variables, such that a high absolute value of the sample correlation coefficient will be considered very suspicious. For example, it will be very difficult to accept a high absolute value of a sample correlation coefficient between two variables that intuitively have nothing in common, i.e., cumulative rainfall in Brazil with inflation rate in Europe or with an index of a stock market in Asia. Contrary, in the case of spurious regressions the analyst will have an indication as to why he or she is getting spurious statistical results. As Granger and Newbold (1974) have pointed out a low value of the Durbin-Watson statistic will appear in regression analysis and standard procedures, such as the Cochrane-Orcutt procedure, will fail to correct the problem of autocorrelated errors. Thus, the analyst must search for other sources that caused the problem of autocorrelation in regression analysis. All of the above findings for spurious behavior can also be found in an interesting paper presented in the literature by Hendry (1980) in which he examines the relationship between cumulative rainfall and inflation rate in UK.

Therefore, it is more difficult to detect spurious correlations as appose to spurious regressions. However, if evidence of spurious behavior exists, this paper demonstrates that using the variance of the sample correlation coefficient of two independent stationary AR(1) processes the spurious behavior is eliminated to a high degree for moderate and large sample sizes. Moreover, a thorough investigation of the spurious correlation phenomenon for two independent non-stationary and stationary AR(1) processes using Monte Carlo analysis is presented in this paper which indicates that the source of this behavior is in the variance of the relevant t statistic.

2. Testing for linear association

To investigate whether or not a linear association between two random variables exists, an analyst will have to test the null hypothesis of zero correlation, i.e., H_0: $\rho = 0$, against the alternative of not zero correlation, i.e., H_1: $\rho \neq 0$, where ρ is the population correlation coefficient. The test is implemented using the following t statistic:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{T-2}}} \tag{1}$$

where r is the sample correlation coefficient and T is the sample size. The t statistic follows a t distribution with (T - 2) degrees of freedom and the null hypothesis will be accepted if its absolute value is less than the critical value, indicating evidence of no linear association between the two variables. In the case of spurious correlations the null hypothesis will be rejected, although the two variables will not be linearly related.

Yule (1926) has indicated that frequently in practice a high degree of linear association between two random variables can arise for variables that have nothing in common, meaning that the typical statistical procedure will support strong evidence of a linear association. As an effort to justify this behavior, Yule (1926) studied the properties of the sample correlation coefficient of two random variables and noticed that the major factor which determines this spurious behavior is the shape of the frequency distribution of the correlation coefficient of the two series. More precisely, if the distribution has a U shape it is certain that spurious correlations will arise.

Banerjee *et al.* (1993), using Monte Carlo analysis, examined the frequency distribution of the correlation coefficient for various orders of integrated independent time series verifying Yule's (1926) initial results. They concluded that if two series are stationary white noise processes, the frequency distribution of the correlation coefficient will be symmetric around zero and it will look like normal distribution. If the two processes are non-stationary *I(1)* processes, the frequency distribution of the correlation coefficient will be semi-ellipse, whereas if the two processes are non-stationary *I(2)* processes, the frequency distribution has a U shape with values of -1 and +1 to be more likely to occur.

Figures 1, 2 and 3 depict these results by presenting the frequency distribution of the sample correlation coefficient of two independent series generated by a) white noise processes, b) non-stationary *I(1)* processes, i.e., $Y_t = Y_{t-1} + u_t$, and c) non-stationary *I(2)* processes, i.e., $Y_t = 2Y_{t-1} - Y_{t-2} + u_t$, for sample size of 100 observations using 10,000 replications respectively.[1] However, there is also another important issue that needs to point out and this issue is related to the distribution of the t statistic for testing the null hypothesis of zero correlation.

Indeed, as can be seen from Table 1, which reports simulation results of spurious correlations for two independent stationary and non-stationary processes, the value of the standard deviation of the t statistic strongly deviates from one for the two non-stationary cases as appose to the white noise case which remains unchanged with value of one, regardless of the sample size. Thus, for two independent *I(1)* processes the null hypothesis at the 5% nominal level will be rejected 76.95% for series of 100 observations, whereas for two independent *I(2)* processes this number becomes 94.82%. Moreover, since the value of the standard deviation of the t statistic increases as the sample size increases, the null hypothesis will be rejected even more frequently, as Table

[1] Note that the shape of these diagrams is not affected by the sample size.

1 reports. It is interesting also to note that for two independent *I(1)* processes the standard deviation of the t statistic ranges from 7.38 to 23.37 for series of 100 to 1,000 observations, whereas for two independent *I(2)* processes this number ranges from 49,01 to 139.39, making the mean value of the t statistic very unreliable. Contrary, no evidence of spurious correlations will appear for the white noise case.

Figures 4, 5 and 6 present the frequency distribution of the t statistic for testing the null hypothesis of zero correlation for series generated by white noise, *I(1)* and *I(2)* processes respectively, for sample size of 100 observations using 10,000 replications.[2] These figures, except for the white noise case, differ significantly from the standard normal distribution. Therefore, the large number of rejections of the null hypothesis of zero correlation which appears for non-stationary processes is basically due to the fact that the variance of the sample correlation coefficient is not properly computed and therefore it affects the value of the t statistic. Especially for two independent non-stationary *I(2)* processes the spurious correlation phenomenon will appear almost with certainty, since the percentage of rejections is almost 100%.

Spurious correlations can also appear in the case of two independent stationary AR(1) processes, as indicated by Granger *et al* (2001) in the context of spurious regressions. To illustrate, consider two independent AR(1) processes X_t and Y_t generated from the following DGP:

$$X_t = \varphi_x X_{t-1} + \varepsilon_{xt} \tag{2}$$

$$Y_t = \varphi_y Y_{t-1} + \varepsilon_{yt} \tag{3}$$

[2] Note that the scale of horizontal axis is not the same for all figures.

where the errors ε_{xt} and ε_{yt} are each white noise N(0, 1) processes independent of each other and the autoregressive parameters are allowed to take values of 0.0, 0.2, 0.5, 0.8 and 0.9. Note that if $\varphi_x = \varphi_y = 1$, both processes are non-stationary random walk processes without drift, whereas if $\varphi_x = \varphi_y = 0$, both processes are white noise processes.

Table 2 reports the percentage of rejections of the null hypothesis of zero correlation against a two sided alternative at the 5% nominal level for two independent stationary AR(1) processes for sample sizes of 100, 500 and 1,000 observations based on 10,000 replications. The most interesting feature of this table is the fact that, unlike the two non-stationary cases previously discussed and especially the *I(1)* case, the percentage of rejections of the null hypothesis of zero correlation remains unchanged regardless of the sample size and it is only affected by the magnitude of the autoregressive parameters. Thus, one will get more spurious results as the value of the autoregressive parameter increases. For example, for $\varphi_x = \varphi_y = 0.5$, the null hypothesis is rejected approximately 13%, for the 5% nominal level, whereas for $\varphi_x = \varphi_y = 0.9$, this number becomes approximately 52%.

Clearly, if the decision, as to whether or not spurious behavior exists, was based on the shape of the frequency distribution of the correlation coefficient, the analyst will have no indication in this case. The frequency distribution for the correlation coefficient of two independent stationary AR(1) processes is symmetric around mean zero and it looks very similar to the white noise case previously presented, as can be seen from Figures 7 and 8 for values of the autoregressive parameters of 0.5 and 0.9 respectively and for sample size of 100 observations. However, as in the case of non-stationary processes, the problem of spurious correlations appears because the value of the standard deviation of the t statistic for testing the null hypothesis of zero correlation is not one for all values of the autoregressive parameters, as Table 3 reports. Thus, although the frequency distribution for the t statistic is symmetric around mean zero, it becomes flatter than the standard normal distribution as the value of the autoregressive parameters increases. Figures 9 and 10, for example, present the frequency distribution for the t statistic of two independent stationary AR(1) processes for values of the autoregressive parameters of 0.5 and 0.9 respectively and for sample size of 100 observations. As can be seen from these figures, the frequency distribution of the t statistic for the 0.9 case is flatter than the 0.5 case. It is also interesting to point out that these estimates of the standard deviation of the t statistic are affected only by the values of the autoregressive parameters and not by the sample size. For example, the standard deviation of the t statistic for $\varphi_x = \varphi_y = 0.5$ is close to 1.3, whereas for $\varphi_x = \varphi_y = 0.9$ this number is approximately equal to 3.

3. Simulation results using the variance of the sample correlation coefficient of two independent stationary AR(1) processes

Having discussed the problem of spurious correlations for two independent stationary AR(1) processes the next step is to investigate the behavior of this phenomenon using the variance of their sample correlation coefficient obtained in the following way. For two independent stationary AR(1) processes X_t and Y_t generated by equations (2) and (3) with autocorrelation coefficients ρ_x and ρ_y respectively we have:

$$E[\frac{1}{T^2}(\sum X_tY_t)^2] = \frac{1}{T^2}E[\sum X_t^2Y_t^2 + 2\sum_{t\neq s} X_tY_tX_sY_s] \tag{4}$$

and since

$$\sum_{t\neq s} X_tY_tX_sY_s = \sum X_tX_{t+1}Y_tY_{t+1} + \sum X_tX_{t+2}Y_tY_{t+2} + \ldots$$

equation (4) is written as:

$$E[\frac{1}{T^2}(\sum X_tY_t)^2] = \frac{1}{T}\sigma_x^2\sigma_y^2 + \frac{2\sigma_x^2\sigma_y^2}{T^2}[(T-1)\rho_x\rho_y + (T-2)\rho_x^2\rho_y^2 + \ldots]$$

which approximately is equal to:

$$E[\frac{1}{T^2}(\sum X_tY_t)^2] = \frac{1}{T}\sigma_x^2\sigma_y^2(1 + \frac{2\rho_x\rho_y}{1-\rho_x\rho_y})$$

Hence, the variance of the sample correlation coefficient between the two independent stationary AR(1) series is approximately defined as:

$$Var(r) = \frac{1}{T}(\frac{1+\rho_x\rho_y}{1-\rho_x\rho_y}) \tag{5}$$

or equivalently as:

$$Var(r) = \frac{1}{T}\left(\frac{1+\phi_x\phi_y}{1-\phi_x\phi_y}\right) \tag{6}$$

since $\rho_x = \varphi_x$ and $\rho_y = \varphi_y$ for AR(1) processes. More evidence about the proof of this variance can be found in Bartlett (1935). In addition, McGregor (1962) verifies the existence of this variance by determining the approximate null distribution of the sample correlation coefficient of two stationary Markov chain processes using the steepest descents method proposed by Daniels (1954 and 1956).

Clearly, the degree of accuracy of this variance depends on the sign, on the absolute magnitude of the two autoregressive coefficients and on the sample size. One should expect less accuracy if φ_x and φ_y are both positive (or negative) and if their absolute magnitude is close to one, as well as if the sample size is small. Therefore, it is interesting to investigate the accuracy of this variance in the context of spurious correlations for all positive values of the autoregressive parameters and for various sample sizes.

For this purpose, series of two independent AR(1) processes X_t and Y_t given by equations (2) and (3) respectively are generated for values of the autoregressive parameter of 0.0, 0.2, 0.5, 0.8 and 0.9 and for sample sizes of 100, 500 and 1,000 observations. Based on the sample correlation coefficient of these two series, the test for zero correlation is conducted by replacing the denominator of the t statistic defined in equation (1) with the square root of the variance of equation (6) and the results of 10,000 replications at the 5% nominal level are reported on Table 4. Perhaps the most astonishing feature of this table is the fact that for almost all cases the empirical levels are very close to nominal levels except for values of the autoregressive parameters of 0.8 and

0.9 and for sample size of 100 observations, in which the test is not working properly. This finding is due to the fact that the frequency distribution of the t statistic for testing the null hypothesis of zero correlation using the variance of two independent AR(1) processes is very close to the standard normal distribution since the standard deviation of the t statistic is one for almost all cases as can be seem from Table 5.

4. Concluding remarks

Spurious relations can be found in practice not only for a pair of two independent non-stationary processes, but also, as Granger *et al* (2001) have indicated, for a pair of two independent stationary AR(1) processes, a result that can also be investigated as a spurious correlation phenomenon. Thus, using the sample correlation coefficient of two independent stationary AR(1) processes, an analyst will find evidence of rejecting the null hypothesis of zero correlation, depending on the value of the autoregressive parameter and not on the sample size.

This study demonstrates that the problem of spurious correlations between two independent stationary AR(1) processes does not come from the frequency distribution of the sample correlation coefficient, as Yule (1926) have suggested and Banerjee *et al.* (1993) have reconfirmed for two independent non-stationary processes, but it comes from the frequency distribution of the relevant t statistic. Therefore, using the approximate variance of the sample correlation coefficient of two independent stationary AR(1) processes, this study shows that the spurious behavior can be eliminated for large and moderate sample sizes, even for large values of the autoregressive parameter.

References

Banerjee, A., Dolado, J., Galbraith, J. W. and Hendry, D. F. (1993). *Co-integration, error correction and the econometric analysis of non-stationary data*, Oxford University Press, Oxford.

Bartlett, M. S. (1935). Some aspects of the time-correlation problems in regard to tests of significance, *Journal of the Royal Statistical Society*, **98**, 536-543.

Daniels, H. E. (1954). Saddlepoint approximations in statistics, *Ann. Math Statistics*, **25**, 631-650.

Daniels, H. E. (1956). The approximate distribution of serial correlation coefficients, *Biometrika*, **43**, 169-185.

Granger, C. W. J. and Newbold, P. (1974). Spurious regressions in econometrics, *Journal of Econometrics*, **2**, 111-120.

Granger, C. W. J., Hyung, N. and Jeon, Y. (2001). Spurious regressions with stationary series, *Applied Economics*, **33**, 899-904.

Hendry, D. F. (1980). Econometrics – Alchemy or science?, *Economica*, **47**, 387-406

McGregor, J. R. (1962). The approximate distribution of the correlation between two stationary linear Markov series, *Biometrika,* **49**, 379-388.

Phillips, P. C. B. (1986). Understanding spurious regressions in econometrics, *Journal of Econometrics*, **33**, 311-340.

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series, *Journal of the Royal Statistical Society*, **89**, 1-64.

Table 1
Simulation results for spurious correlations with white noise and non-stationary I(1) and I(2) processes based on 10,000 replications

T	Percentage of rejections of H_0 ($\|t\| > 1.96$)	Mean value of r	Standard deviation of r	Mean value of t	Standard deviation of t
WHITE NOISE PROCESSES					
100	5.03	0.00	0.10	0.00	1.00
500	5.24	0.00	0.04	0.00	1.00
1,000	5.10	0.00	0.03	0.01	1.00
***I(1)* PROCESSES**					
100	76.95	0.00	0.49	0.03	7.38
500	89.32	0.00	0.49	0.01	16.37
1,000	92.76	0.00	0.49	0.07	23.37
***I(2)* PROCESSES**					
100	94.82	0.00	0.84	0.03	49.01
500	97.51	0.00	0.82	0.28	98.55
1,000	98.18	-0.01	0.82	-0.40	139.39

Table 2
Percentage of rejections of the null hypothesis of zero correlation at the 5% nominal level (|t| > 1.96) for two independent stationary AR(1) processes based on 10,000 replications

		φ_x				
T	φ_y	**0.0**	**0.2**	**0.5**	**0.8**	**0.9**
100	**0.0**	5.03				
	0.2	5.18	6.20			
	0.5	5.38	7.77	12.85		
	0.8	5.04	9.86	19.53	35.23	
	0.9	4.81	10.44	22.42	41.12	50.27
500	**0.0**	5.24				
	0.2	5.13	5.77			
	0.5	5.31	7.92	13.15		
	0.8	4.90	9.44	19.72	34.78	
	0.9	5.42	10.46	22.67	42.56	52.22
1,000	**0.0**	5.10				
	0.2	5.25	6.00			
	0.5	4.73	7.38	12.87		
	0.8	5.30	9.94	19.92	36.05	
	0.9	4.83	10.53	22.55	43.21	52.01

Table 3
Standard deviation of the t statistic for testing the null hypothesis of zero correlation for two independent stationary AR(1) processes based on 10,000 replications

T	φ_y	φ_x 0.0	0.2	0.5	0.8	0.9
100	**0.0**	1.00				
	0.2	1.02	1.05			
	0.5	1.01	1.11	1.30		
	0.8	1.00	1.17	1.52	2.12	
	0.9	1.02	1.21	1.62	2.45	3.00
500	**0.0**	1.00				
	0.2	1.01	1.04			
	0.5	1.00	1.11	1.29		
	0.8	1.02	1.18	1.51	2.10	
	0.9	1.00	1.21	1.62	2.48	3.10
1,000	**0.0**	1.00				
	0.2	1.00	1.05			
	0.5	1.00	1.11	1.28		
	0.8	1.01	1.17	1.52	2.15	
	0.9	1.00	1.20	1.61	2.47	3.07

Table 4
Percentage of rejections of the null hypothesis of zero correlation at the 5% nominal level (|t| > 1.96) for two independent stationary AR(1) processes using the approximate variance of their sample correlation coefficient based on 10,000 replications

T	φ_y	φ_x 0.0	0.2	0.5	0.8	0.9
100	**0.0**	4.76				
	0.2	5.22	5.13			
	0.5	5.18	5.08	4.56		
	0.8	4.96	4.72	4.59	3.41	
	0.9	5.08	4.73	4.46	2.75	1.59
500	**0.0**	5.21				
	0.2	5.09	5.03			
	0.5	5.07	5.03	5.25		
	0.8	4.91	4.77	5.07	4.81	
	0.9	5.09	4.88	5.11	4.61	4.56
1,000	**0.0**	4.98				
	0.2	5.21	4.81			
	0.5	4.90	5.13	4.90		
	0.8	4.80	5.08	5.09	4.77	
	0.9	5.09	4.88	4.92	4.79	5.16

Table 5
Standard deviation of the t statistic for testing the null hypothesis of zero correlation for two independent stationary AR(1) processes using the approximate variance of their sample correlation coefficient based on 10,000 replications

T	φ_y	φ_x				
		0.0	**0.2**	**0.5**	**0.8**	**0.9**
100	**0.0**	1.01				
	0.2	1.01	1.00			
	0.5	1.00	1.00	0.98		
	0.8	1.01	1.00	0.99	0.95	
	0.9	1.01	1.00	0.98	0.92	0.88
500	**0.0**	1.00				
	0.2	1.00	1.00			
	0.5	1.01	1.01	1.01		
	0.8	1.00	0.99	0.99	0.99	
	0.9	1.01	1.00	1.00	0.99	0.99
1,000	**0.0**	1.00				
	0.2	1.00	1.00			
	0.5	1.00	1.01	1.00		
	0.8	0.99	1.01	1.01	0.99	
	0.9	1.01	1.00	1.00	1.00	1.00

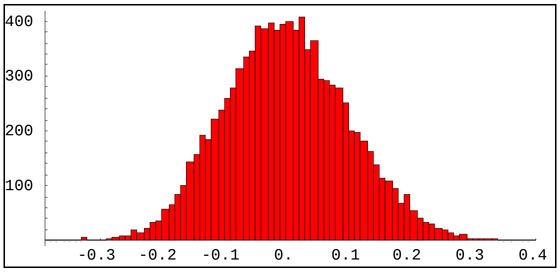

Figure 1
Frequency distribution for the correlation coefficient between two independent white noise processes (*T*=100)

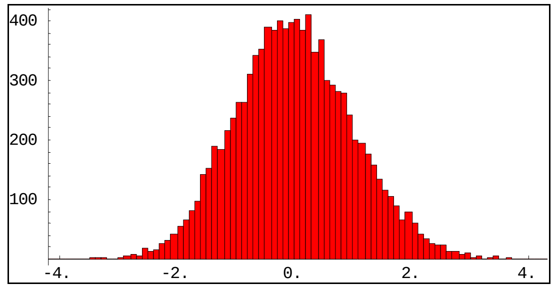

Figure 4
Frequency distribution for the t statistic between two independent white noise processes (*T*=100)

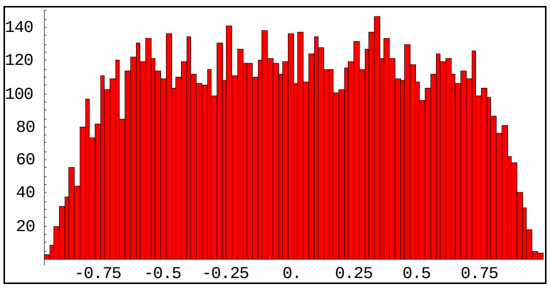

Figure 2
Frequency distribution for the correlation coefficient between two independent *I*(1) processes (*T*=100)

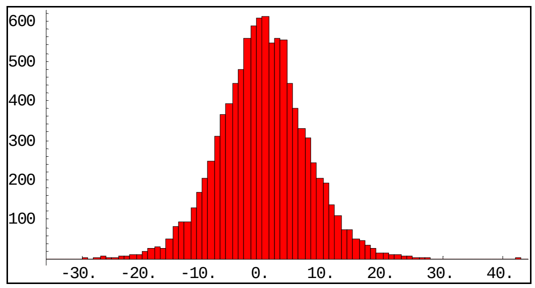

Figure 5
Frequency distribution for the t statistic between two independent *I*(1) processes (*T*=100)

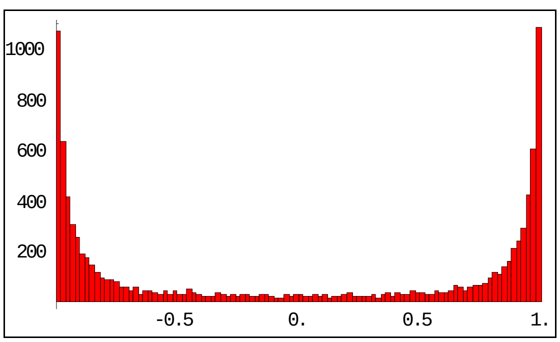

Figure 3
Frequency distribution for the correlation coefficient between two independent *I*(2) processes (*T*=100)

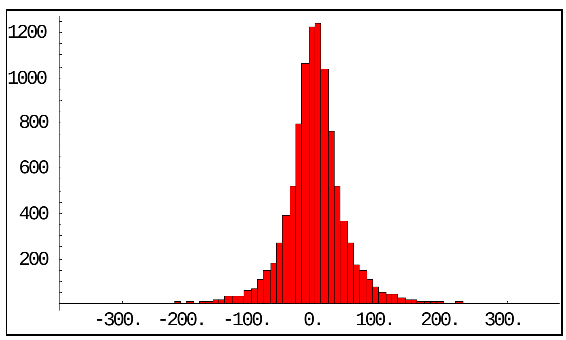

Figure 6
Frequency distribution for the t statistic between two independent *I*(2) processes (*T*=100)

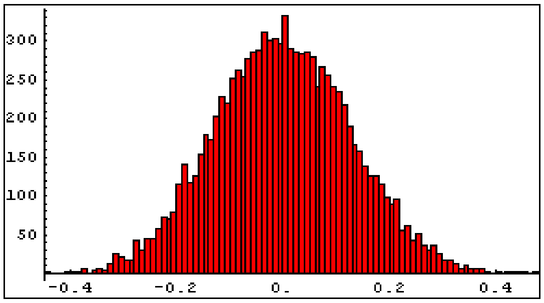

Figure 7

Frequency distribution for the correlation coefficient between two independent AR(1) processes for $\varphi_\chi = \varphi_y = 0.5$ (T=100)

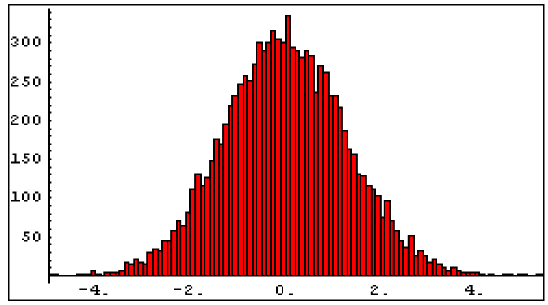

Figure 9

Frequency distribution for the t statistic between two independent AR(1) processes for $\varphi_\chi = \varphi_y = 0.5$ (T=100)

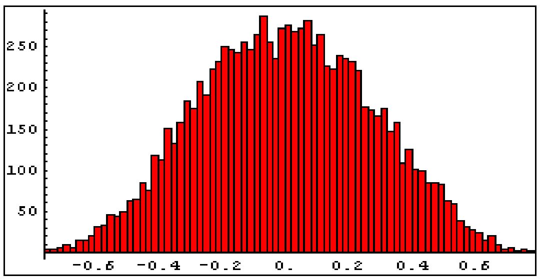

Figure 8

Frequency distribution for the correlation coefficient between two independent AR(1) processes for $\varphi_\chi = \varphi_y = 0.9$ (T=100)

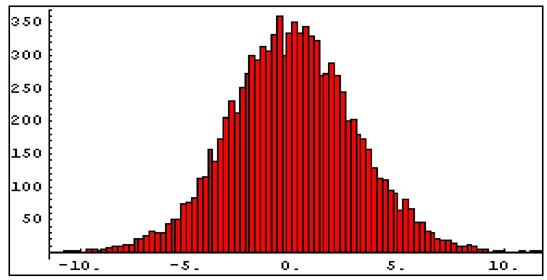

Figure 10

Frequency distribution for the t statistic between two independent AR(1) processes for $\varphi_\chi = \varphi_y = 0.9$ (T=100)