15381 NATIONAL LIBRARY
OTTAWA

BIBLIOTHÈQUE NATIONALE
OTTAWA

NAME OF AUTHOR....MARGARET L.C. WOO.....

TITLE OF THESIS....A SIMULATION STUDY OF THE DISTRIBUTION
....OF CORRELATION BETWEEN TWO LINEAR
....STATIONARY MARKOV PROCESSES.....

UNIVERSITY...UNIVERSITY OF ALBERTA.....

DEGREE FOR WHICH THESIS WAS PRESENTED. MASTER OF SCIENCE

YEAR THIS DEGREE GRANTED.....1973.....

Permission is hereby granted to THE NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed)...Margaret Woo.....

PERMANENT ADDRESS:

15-B MINBU ROAD,.....
MANDALAY MANSIONS,.....
SINGAPORE 11, SINGAPORE.

DATED....12th JAN.....1973

NL-91 (10-68)

THE UNIVERSITY OF ALBERTA

A SIMULATION STUDY OF THE DISTRIBUTION OF CORRELATION BETWEEN TWO LINEAR STATIONARY MARKOV PROCESSES

by

(C) MARGARET L. C. WOO

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

SPRING, 1973

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled A SIMULATION STUDY OF THE DISTRIBUTION OF CORRELATION BETWEEN TWO LINEAR STATIONARY MARKOV PROCESSES, submitted by Margaret L.C. Woo in partial fulfillment of the requirements for the degree of Master of Science.

Supervisor

Date Dec. 18, 1972.

ABSTRACT

This thesis discusses the application of a simulation method to determine the accuracy of an approximate distribution of the sample cross-correlation between two linear, stationary Markov series with known autocorrelations of lag one, ρ_1 and ρ_2 . The topics discussed are: the simulation of the sample cross-correlation distribution; the generation of pseudo-random numbers; the statistical testing of simulation results; the determination of the critical values of the approximate distribution; the application of the simulated distribution to estimate the accuracy in the approximate distribution and its critical values.

The results presented here show that the approximate distribution is accurate for low values of the product of autocorrelations $\rho_1\rho_2$ ($|\rho_1\rho_2| \leq .5$). For high values of the autocorrelations and small sample sizes (≤ 30) of the Markov series, however, the approximate distribution appears to have too large a variance.

A practical example is used to illustrate how the approximate distribution may be applied to test for correlation between two series of data of known autocorrelations.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to my supervisor Professor U.M. von Maydell for her constant guidance, encouragement, patience and financial support at all stages of this research. This work was supported partly by the National Research Council of Canada.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

THE APPROXIMATE NULL DISTRIBUTION OF THE SAMPLE CROSS-CORRELATION BETWEEN TWO LINEAR STATIONARY MARKOV SERIES.

1.1 Introduction.

The testing for correlation between two variables is often desired, but well-known tests for this purpose are generally based on the assumption that at least one variable is not autocorrelated (that is, there is no interdependence or serial correlation between successive observations of the variable). In economic, meteorological, biological and some other time series, autocorrelation usually exists. Hence, there is a real need for a test for correlation between two autocorrelated series.

To use hypothesis testing, the null distribution of the sample correlation must be available. For two series of independent observations, the sample cross-correlation r is known to have the null distribution of the Pearson correlation coefficient (see Keeping [21]). However, if both series are autocorrelated, very little is known of the exact distribution of r.

Since the mathematical model of a time series is a stochastic process, the most promising approach to the

satisfactory analysis of time series is the use of stochastic processes. McGregor and Bielenstein [30] have derived an approximate null distribution of the cross-correlation between two autocorrelated series which are generated by stationary, linear Markov processes. This approximate distribution depends only on the size of the sample taken in each series and the product of the autocorrelations in the series.

In order to apply the approximate distribution of r in any valid test for correlation, it is necessary to determine the accuracy of this distribution and to compute the critical values of r. This paper will consider; a) the use of simulation to check the accuracy of the approximate distribution; b) an algorithm for evaluating the critical values of r; c) the estimation of error bounds for the approximate critical values of r; and d) the application of the approximate distribution and critical values in some practical examples.

An attempt at obtaining the accuracy of the approximate distribution is made by comparing this distribution to a simulated distribution of the cross-correlation coefficient. Simulation of the cross-correlation distribution requires the use of a large number of normal random numbers. The simulation procedure is described in Chapter 2. Chapter 3 will discuss the properties and choice of a pseudo-random number generator suitable for use in the simulation. Before the simulated distribution can be used for any purpose, it is necessary to

determine how closely it represents the true distribution. Chapter 4 discusses a series of test criteria which can be used to determine the goodness-of-fit of the simulated distribution. The comparison of the approximate and simulated distributions, and the estimation of the error in the approximate critical values of r are discussed in Chapter 5. Chapter 6 will consist of an example of application of the approximate distribution, the conclusion, and discussions of further research to improve the simulation efficiency.

1.2 The Approximate Null Distribution of the Sample Correlation between Two Linear, Stationary Markov Processes.

Let (x_t , y_t), $t=1,2,\ldots,n$, be a sample of n pairs of values observed from two linear, stationary Markov processes (that is, first order autoregressive processes) defined by:

$$\begin{aligned} X_t &= \rho_1 X_{t-1} + Z_t \\ Y_t &= \rho_2 Y_{t-1} + Z'_t \end{aligned} \qquad (1.1)$$

where Z_t and Z'_t are assumed to be independent, normal, N(0,1) random variables; and the autocorrelations of lag one of the processes, ρ_1 and ρ_2, respectively are assumed to be known.

The cross-correlation (or product correlation) between the two stationary processes, X_t , Y_t is defined by:

$$\rho_{XY} = \frac{\mathrm{Cov}(X_t , Y_t)}{[\mathrm{Var}(X_t) \cdot \mathrm{Var}(Y_t)]^{1/2}} \tag{1.2}$$

it is a number between -1 and +1, and is zero when X_t and Y_t are uncorrelated.

An estimator of ρ_{XY} is the sample cross-correlation coefficient r_{XY} , given by:

$$r_{XY} = \frac{\sum_{t=1}^{n} (X_t - \bar{X})(Y_t - \bar{Y})}{[\sum_{t=1}^{n}(X_t - \bar{X})^2 \sum_{t=1}^{n}(Y_t - \bar{Y})^2]^{1/2}} \tag{1.3}$$

where

$$\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t \quad \text{and} \quad \bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t \quad .$$

The random variable r_{XY} is an unbiased estimator of ρ_{XY} if $\rho_{XY} = 0$, since $E(r_{XY}) = \rho_{XY}$.

Processes of the type given in Eqn. (1.1) are frequently used as sampling models in the studies of economic time series; and testing for correlation between two series generated by these processes is often desired. It is often required to test the null hypothesis that the two processes X_t , Y_t are

uncorrelated (that is, $\rho_{XY} = 0$). For this purpose the distribution of the sample cross-correlation r_{XY}, under the assumption $\rho_{XY} = 0$, must be available, as well as the corresponding and appropriate critical values for r_{XY}.

When either $\{x_t\}$ or $\{y_t\}$ or both, for t=1,2,...,n, are series of independent observations (that is, either ρ_1 or ρ_2 or both are zero so that at least one of the series is a sample from a normal distribution) the sample cross-correlation r_{XY} has the following null distribution of the Pearson correlation coefficient (see Keeping [21]),

$$p(r) = \frac{[\ 1 - r^2\ ]^{(n-4)/2}}{B[\ \frac{1}{2}\ (n-2)\ ,\ \frac{1}{2}\ ]} \qquad (1.4)$$

where r is a value of the random variable r_{XY}, p(r) is the probability density function of r_{XY}, n is the number of observations taken in each Markov series, and B is the Beta function. In this case the two processes X_t, Y_t are uncorrelated, $\rho_{XY} = 0$. The density function p(r) is a symmetrical, bell-shaped curve for $n \geq 5$ so that

$$E(r_{XY}) = 0 .$$

For an illustration of the graph of p(r) see Fig. 1.1 with $\rho_1\rho_2 = 0$. It can be shown that

$$E(r_{XY}^2) = (\ n - 1\ )^{-1}$$

and hence

$$Var(r_{XY}) = (\ n - 1)^{-1} .$$

The kurtosis is $-6/(n-1)$, which tends to zero as n increases. For very large n, the distribution is approximately normal.

When both series $\{x_t\}$, $\{y_t\}$ are autocorrelated, the Pearson distribution no longer holds and very little is known of the exact distribution of the sample correlation r_{XY} .

McGregor and Bielenstein [30] have found an approximate probabilty density function of the sample correlation r_{XY} under the hypothesis that the population cross-correlation is zero, that is, $\rho_{XY} = 0$. For a random sample $\{ (x_t , y_t) , t=1,2,\ldots,n \}$ generated by the two processes in Eqn. (1.1) the approximate density function, $p^*(r)$, of r_{XY} was derived to be

$$p^*(r) = \frac{2^{M-3}(1-\rho_1\rho_2)^{1/2}}{B[\frac{1}{2}M-1, \frac{1}{2}]} \times \frac{(1-r^2)^{(M-4)/2}}{\{[(1+\rho_1\rho_2)^2 - 4\rho_1\rho_2 r^2]^{\frac{1}{2}} + (1-\rho_1\rho_2)\}^{M-\frac{3}{2}}}$$

$$\times \frac{\{[(1+\rho_1\rho_2)^2 - 4\rho_1\rho_2 r^2]^{\frac{1}{2}} + (1+\rho_1\rho_2)\}^{\frac{1}{2}}}{[(1+\rho_1\rho_2)^2 - 4\rho_1\rho_2 r^2]^{1/2}} \times \{1 + O(\frac{1}{n})\}, \tag{1.5}$$

where

$$M = n + \frac{\rho_1\rho_2(6 - 5\rho_1\rho_2)}{1 - (\rho_1\rho_2)^2} > 2 ,$$

$$|\rho_1| < 1 , \quad |\rho_2| < 1 , \quad |r| \leq 1 , \quad n \geq 6 ,$$

and

$$B[\tfrac{1}{2}M - 1 , \tfrac{1}{2}] = \frac{\Gamma(\frac{1}{2}M - 1)\ \Gamma(\frac{1}{2})}{\Gamma(\frac{1}{2}M - \frac{1}{2})} .$$

The derived approximate density function, $p^*(r)$, depends only on the sample size n and the product of the autocorrelations, $\rho_1\rho_2$. For the special case when the product of the autocorrelations is zero (that is, $\rho_1\rho_2 = 0$) the approximate density function reduces to the null distribution p(r) given in Eqn. (1.4).

The k^{th} moment, m_k^* , of the approximate distribution may be evaluated by the following general formula :

$$m_k^* = \int_{-1}^{1} r^k . p^*(r)\ dr .$$

The variance, skewness and kurtosis are given by :

$$\text{Variance} = m_2^* - (m_1^*)^2$$

$$\text{Skewness} = m_3^* / (m_2^*)^{3/2}$$

$$\text{Kurtosis} = m_4^* / (m_2^*)^2$$

Computation of the above statistics involves numerical integration and evaluation of $p^*(r)$. The calculation procedure used is similar to that described in Appendix A for the evaluation of the critical values of r_{XY}. The above statistics for the $p^*(r)$ distribution for n = 10, with $\rho_1\rho_2 = -.25, 0, .5$ and n = 15, 30, 50, 75, 100, with $\rho_1\rho_2 = -.5, 0, .5$ are tabulated in Table 1.1. Plots of $p^*(r)$ for these values of $(n, \rho_1\rho_2)$ are shown in Figs. 1.1 (a), (b), (c), where the same scale is used for all three figures. The graphs show that the variance of the distributions for $\rho_1\rho_2 > 0$ is larger than that of the $\rho_1\rho_2 = 0$ distribution and the variance of r_{XY} for $\rho_1\rho_2 < 0$ is less than that for $\rho_1\rho_2 = 0$.

Table 1.1

Moments, Skewness and Kurtosis of Approximate Distribution

n	$\rho_1\rho_2$	Mean	Variance	3rd Mom.	4th Mom.	Skewness	Kurtosis
10	-.25	0.0	.1002	0.0	.0267	0.0	2.663
	.0	0.0	.1111	0.0	.0303	0.0	2.450
	.5	0.0	.1956	0.0	.0782	0.0	2.043
15	-.5	0.0	.0550	0.0	.0097	0.0	3.216
	.0	0.0	.0714	0.0	.0134	0.0	2.625
	.5	0.0	.1460	0.0	.0469	0.0	2.205
30	-.5	0.0	.0162	0.0	.0084	0.0	3.200
	.0	0.0	.0345	0.0	.0033	0.0	2.807
	.5	0.0	.0837	0.0	.0173	0.0	2.464
50	-.5	0.0	.0082	0.0	.0002	0.0	3.122
	.0	0.0	.0204	0.0	.0012	0.0	2.882
	.5	0.0	.0536	0.0	.0075	0.0	2.623
75	-.5	0.0	.0051	0.0	.0001	0.0	3.046
	.0	0.0	.0135	0.0	.0053	0.0	2.921
	.5	0.0	.0370	0.0	.0037	0.0	2.724
100	-.5	0.0	.0037	0.0	.0000	0.0	2.834
	.0	0.0	.0101	0.0	.0003	0.0	2.941
	.5	0.0	.0283	0.0	.0022	0.0	2.782

Fig.1.1(a)

Approximate Density Function of Cross-Correlation Coefficient r.

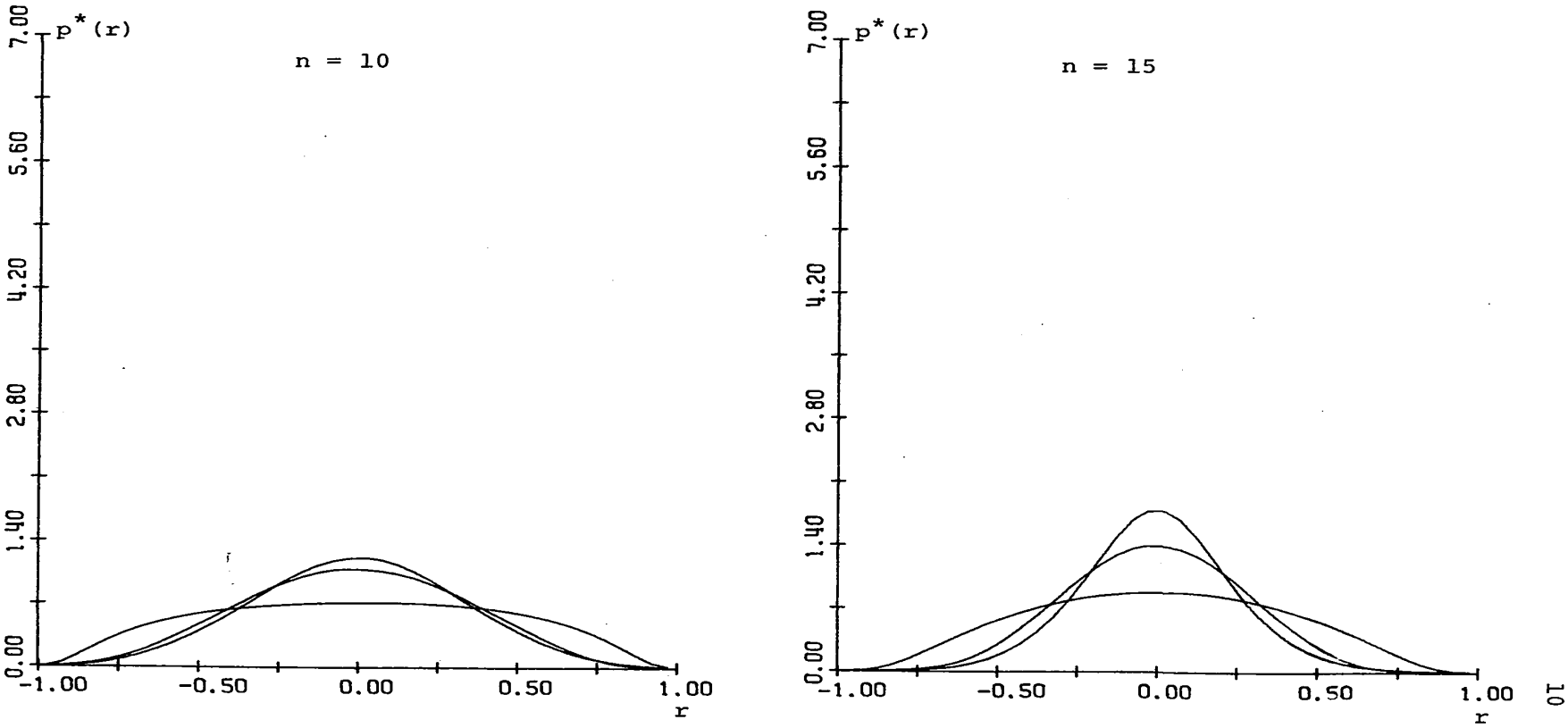
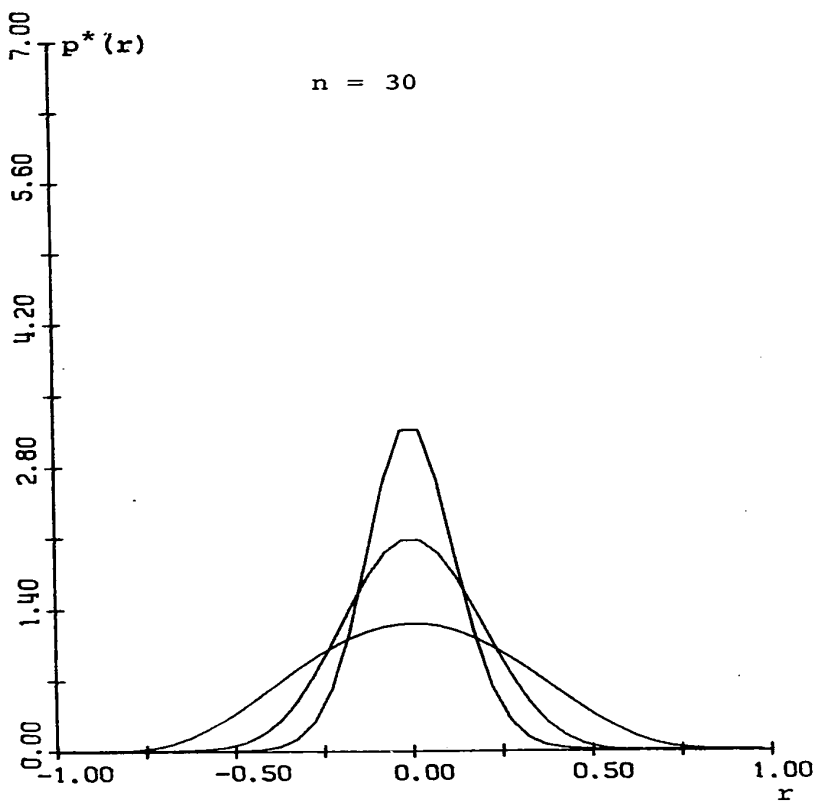


Fig.1.1(b)

Approximate Density Function of Cross-Correlation Coefficient r.



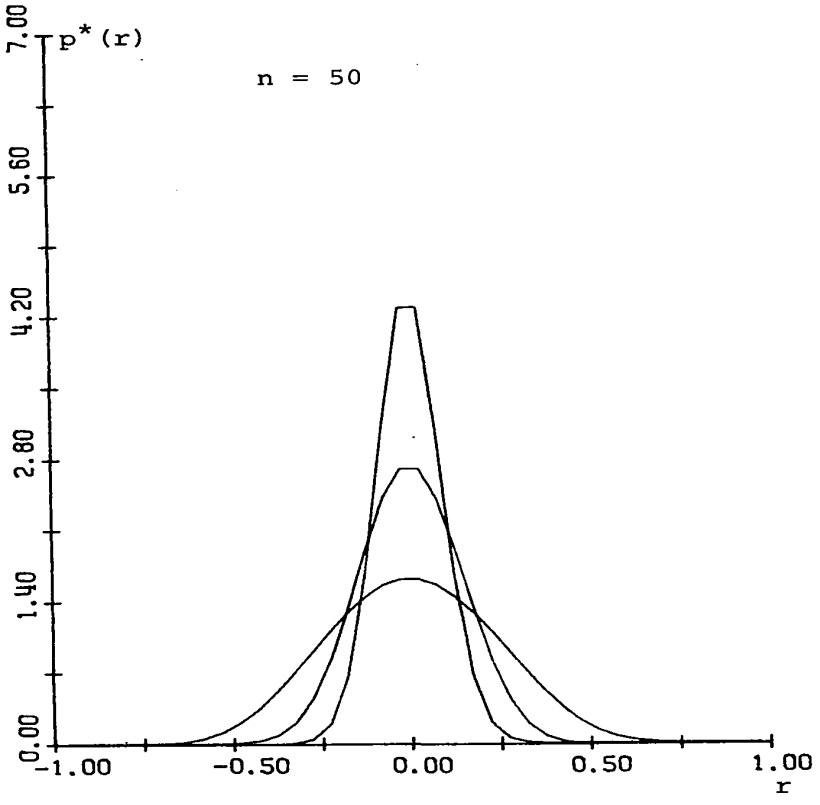
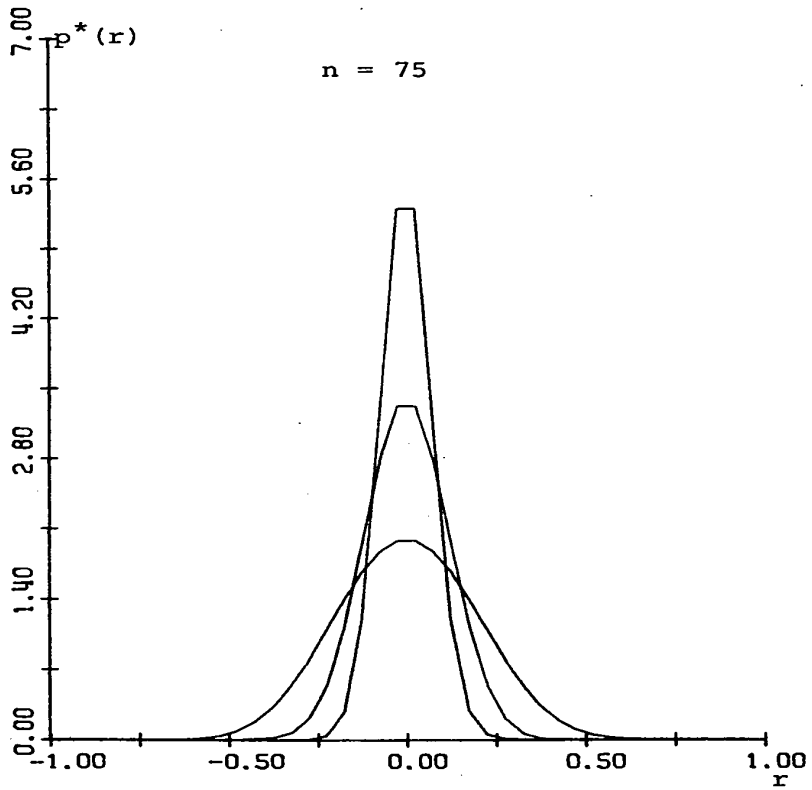
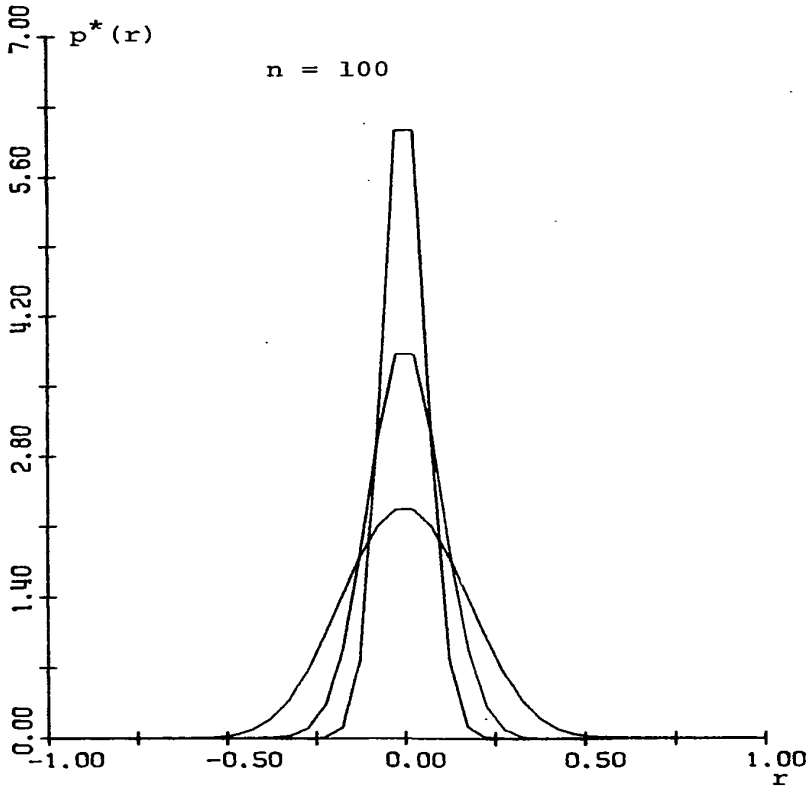


Fig.1.1(c)

Approximate Density Function of Cross-Correlation Coefficient r.





1.3 Outlining the Steps used in Studying the Approximate Distribution of the Sample Cross-Correlation.

In order to apply the approximate density function of r_{XY} in any valid test for correlation the following points have to be considered :

a) Since the derived density function of r_{XY} (Eqn.(1.5)) is only an approximation dependent on the two parameters, the sample size n and the product of the autocorrelations, $\rho_1\rho_2$, it is necessary to determine the accuracy of the approximate distribution for various combinations of these parameters, particularly for small sample sizes (since the approximation is of $O(\frac{1}{n})$ and partly for the reason stated in (b) below).

b) It is also necessary to establish a minimum sample size for which the approximate density function may be used with acceptable accuracy, since in practice the size of series of data or items available for analysis is usually small.

c) In order to test the null hypothesis that two autocorrelated series are uncorrelated, critical values of r_{XY} for appropriate values of $\rho_1\rho_2$, under the assumption that $\rho_{XY} = 0$, have to be available. Hence, these critical

values have to be computed from the approximate density function for each suitable value of $\rho_1\rho_2$.

These points of interest will form the subject of discussion of this thesis. The thesis will focus on :

1) The design of an algorithm for evaluating the critical values of r_{XY} using the approximate density function. Discussion of the algorithm, including numerical integration, will be contained in Appendix A.

2) The simulation of linear stationary Markov series and the distribution of the cross-correlation coefficient r_{XY} in Chapter 2. The discussion will cover a method of simulating realizations of linear stationary Markov processes, including a criterion for determining stationarity of the processes, and of obtaining an empirical distribution of the sample cross-correlation of these realizations.

3) The generation of pseudo-random numbers, including techniques of generating normal random numbers and criteria for testing the quality of a random number generator in Chapter 3. This chapter will also consider the desired properties and the choice of a suitable generator for use in the simulation.

4) The determination of the error in simulation by a series of goodness-of-fit tests and the estimation of a suitable sample size to use for a desired accuracy in the simulation in Chapter 4.

5) The use of simulation as a method for determining the accuracy of the approximate distribution of r_{XY}, especially for small sample sizes in Chapter 5. This chapter will study methods of comparing the simulated and approximate distributions and critical values.

6) The estimation of error bounds for critical values of r_{XY} as computed from the approximate density function in Section 5.4.

7) The application of the approximate distribution in a practical example in Chapter 6.

CHAPTER II

SIMULATION OF THE CROSS-CORRELATION DISTRIBUTION.

2.1 Simulation of Linear Stationary Markov Processes.

In order to study the distribution of cross-correlation between two linear, stationary Markov series for a variety of values of sample sizes and autocorrelations, only simulation provides ready access to a large number of these cross-correlations.

To simulate realizations of a stationary process of the form,

$$x_t = \rho x_{t-1} + z_t \qquad (2.1)$$

the following requirements are evident :

a) an initial starting value, x_0 ;

b) a 'good' random number generator which will produce independent normal random numbers ;

c) a criterion for determining the stage at which the process may be considered to have become stationary.

The starting value x_0 may be arbitrarily chosen as it will be shown that the behaviour of the stationary Markov process is independent of x_0 .

Chapter 3 will discuss the choice of a 'good' random number generator and will contain a description of the random generator that will be used in the simulation.

The criterion for determining stationarity of the process in Eqn.(2.1) will be discussed in the following section.

2.2 Determination of Stationarity of Simulated Markov Process.

A stochastic process X_t is said to be stationary up to order K if and only if ,

1) the mean and variance are constants, independent of t ;

2) the covariance $Cov(X_s , X_t)$ is a function of the lag $|s - t|$ only;

3) all moments of the form

$$E(X_{t_1}^{k_1} , X_{t_2}^{k_2} , \ldots , X_{t_n}^{k_n})$$

up to order

$$k_1 + k_2 + \ldots + k_n = K ,$$

where k_i and K are integers, depend only on the time lag $|(t_i+u) - t_i|$.

Consider the first order autoregressive process defined by

$$X_t = \rho X_{t-1} + Z_t \tag{2.2}$$

where ρ is the autocorrelation of lag one of the process and $\{Z_t\}$ is a sequence of independent random variables of known constant mean μ_z and constant variance σ_z^2. By successive substitution, Eqn. (2.2) may be rewritten as

$$X_t = \rho^t X_0 + \sum_{i=0}^{t-1} \rho^i Z_{t-i} \tag{2.3}$$

where X_0 is the initial value of the process. Taking expectation, we have

$$\begin{aligned} E(X_t) &= \rho^t X_0 + \sum_{i=0}^{t-1} \rho^i E(Z_{t-i}) \\ &= \rho^t X_0 + \mu_z (1 - \rho^t)/(1 - \rho) \\ &= [X_0 - \mu_z/(1-\rho)]\rho^t + \mu_z/(1-\rho), \end{aligned} \tag{2.4}$$

since $E(Z_{t-i}) = \mu_Z$ and $|\rho| < 1$. The variance of the process is given by

$$\begin{aligned}
\mathrm{Var}(X_t) &= E\{[X_t - E(X_t)]^2\} \\
&= E\{[\sum_{i=0}^{t-1}\rho^i Z_{t-i} - \sum_{i=0}^{t-1}\rho^i \mu_Z]^2\} \\
&= E\{[\sum_{i=0}^{t-1}\rho^i (Z_{t-i} - \mu_Z)]^2\} \\
&= E\{\sum_{i=0}^{t-1}[\rho^{2i}(Z_{t-i} - \mu_Z)^2 \\
&\quad + 2\rho^i[(Z_t - \mu_Z)(Z_{t-1} - \mu_Z) \\
&\quad + \ldots + (Z_2 - \mu_Z)(Z_1 - \mu_Z)]]\} \\
&= \sum_{i=0}^{t-1}\{\rho^{2i}E[(Z_{t-i} - \mu_Z)^2] \\
&\quad + 2\rho^i[E(Z_t - \mu_Z)E(Z_{t-1} - \mu_Z) \\
&\quad + \ldots + E(Z_2 - \mu_Z)E(Z_1 - \mu_Z)]\} \\
&= \sum_{i=0}^{t-1}\{\rho^{2i}\sigma_Z^2 \\
&\quad + 2\rho^i[(E(Z_t) - \mu_Z)(E(Z_{t-1}) - \mu_Z) \\
&\quad + \ldots + (E(Z_2) - \mu_Z)(E(Z_1) - \mu_Z)]\} \\
&= \sigma_Z^2 (1 - \rho^{2t}) / (1 - \rho^2) , \qquad (2.5)
\end{aligned}$$

since $\{Z_t\}$ are independent random variables of constant mean and variance. The covariance of X_s, X_t, $s \geq t$, may be expressed as follows :

$$\mathrm{Cov}(X_t , X_s) = E\{[X_t - E(X_t)][X_s - E(X_s)]\}$$

$$= E\{[\sum_{i=0}^{t-1}\rho^i (Z_{t-i} - \mu_z)][\sum_{j=0}^{s-1}\rho^j (Z_{s-j} - \mu_z)]\}$$

$$= E\{[\sum_{i=0}^{t-1}\rho^i (Z_{t-i} - \mu_z)][\sum_{j=0}^{(s-t)-1}\rho^j (Z_{s-j} - \mu_z) + \rho^{s-t}\sum_{i=0}^{t-1}\rho^i (Z_{t-i} - \mu_z)]\}$$

$$= E\{\rho^{s-t}[\sum_{i=0}^{t-1}\rho^i (Z_{t-i} - \mu_z)]^2 + [\sum_{i=0}^{t-1}\rho^i (Z_{t-i} - \mu_z)][\sum_{j=0}^{(s-t)-1}\rho^j (Z_{s-j} - \mu_z)]\}$$

$$= \rho^{s-t}E\{[\sum_{i=0}^{t-1}\rho^i (Z_{t-i} - \mu_z)]^2\} + [\sum_{i=0}^{t-1}\rho^i E(Z_{t-i} - \mu_z)][\sum_{j=0}^{(s-t)-1}\rho^j E(Z_{s-j} - \mu_z)]$$

$$= \sigma_z^2 \rho^{s-t}(1 - \rho^{2t}) / (1 - \rho^2) , \qquad (2.6)$$

using the result obtained in Eqn. (2.5).

For the process in Eqn. (2.2) to be stationary, the mean, $E(X_t)$, and variance, $\mathrm{Var}(X_t)$, must be independent of t, and the covariance, $\mathrm{Cov}(X_t , X_s)$, must depend only on the lag $|s-t|$. Since $|\rho| < 1$,

$$\lim_{t\to\infty} \rho^t = 0$$

and from Eqns. (2.4) , (2.5) and (2.6) , we have as t tends to infinity,

$$\lim_{t\to\infty} [E(X_t)] = \mu_z / (1 - \rho) ,$$

$$\lim_{t\to\infty} [Var(X_t)] = \sigma_z^2 / (1 - \rho^2) ,$$

$$\lim_{t\to\infty} [Cov(X_t , X_s)] = \sigma_z^2 \rho^{s-t} / (1 - \rho^2) .$$

Hence, the process (2.2) is stationary only after a sufficient number of terms has been generated. The time point at which the process may be considered to have stabilized can be determined by setting ρ^t equal to a very small value (that is, a numerical zero). In the simulation the following condition for stationarity is used :

$$\rho^t = 10^{-8} . \qquad (2.7)$$

Hence, the time point t_x after which the process X_t stabilizes is given by

$$t_x = \ln 10^{-8} / \ln | \rho | . \qquad (2.8)$$

At time point t_x we may write

$$E(X_t) = \mu_z / (1 - \rho) + \varepsilon$$

$$Var(X_t) = \sigma_z^2 / (1 - \rho^2) + \varepsilon'$$

$$\text{Cov}(X_t , X_s) = \sigma_z^2 \rho^{s-t} / (1 - \rho^2) + \varepsilon''$$

where ε , ε' , ε'' are small error terms (of the order of 10^{-8}). Since the errors are small we may assume that at this time point we have a stationary series.

As can be seen from Eqn. (2.5) and (2.6) , the variance and covariance are independent of the initial value X_0 ; and by setting ρ^t equal to a numerical zero as a condition for stationarity the mean becomes independent of X_0 . Hence, in the stationary state the process X_t does not depend on the starting value X_0 .

For the case where $\{Z_t\}$ is a sequence of independent $N(0, 1)$ random variables we have as t tends to infinity,

$$\lim_{t\to\infty} [E(X_t)] = 0$$

$$\lim_{t\to\infty} [\text{Var}(X_t)] = 1 / (1 - \rho^2)$$

$$\lim_{t\to\infty} [\text{Cov}(X_t , X_s)] = \rho^{s-t} / (1 - \rho^2)$$

2.3 Generation of the Cross-Correlation Frequency Distribution.

To obtain empirical distributions of the cross-correlations between autocorrelated series, a set of observations of two series is generated by means of processes of the form in Eqn.(2.2) and the sample cross-correlation between these series is computed using the formula given below. The following notation is used :

X_t , Y_t - linear, stationary Markov processes defined by

$$X_t = \rho_1 X_{t-1} + Z_t$$

$$Y_t = \rho_2 Y_{t-1} + Z'_t \qquad (2.9)$$

ρ_1 , ρ_2 - known autocorrelations of lag one

Z_t , Z'_t - independent normal N(0,1) random variables

t_x , t_y - time-points at which X_t , Y_t , respectively, reach stationarity as determined by Eqn. (2.8)

n - the number of realizations of each of the processes X_t , Y_t, that is, the sample size or length of the time series

r_{XY} - the sample cross-correlation between the X_t and Y_t processes

r - a value of the random variable r_{XY}

N - the number of simulated observations r of r_{XY}.

The following procedure is used to simulate N values of the sample cross-correlation between two series generated by X_t, Y_t :

1) The first t_x sucessive values of $\{x_t\}$ are generated and discarded.

2) The next n generated values are then taken to be the sample observations of the $\{x_t\}$ series.

3) The above process is repeated to obtain a sample of n observations of the $\{y_t\}$ series.

4) The values of the sample cross-correlation r_{XY} for this particular (x_t, y_t) sample is then evaluated by the following formula :

$$r = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})}{\left[\sum_{t=1}^{n} (x_t - \bar{x})^2 \sum_{t=1}^{n} (y_t - \bar{y})^2\right]^{1/2}}$$

where

$$\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t \quad , \quad \bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t .$$

5) Steps (2) , (3) and (4) are repeated N times to obtain N simulated values of the sample cross-correlation coefficient r_{XY} between the $\{x_t\}$ and $\{y_t\}$ series.

The sample of N observations of r generated by this procedure is then organized and summarized to obtain the following :

1) A frequency table as shown in Tables 4.2 and 4.3.

2) A graph of the relative frequency distribution, representing the empirical probability density function of r_{XY}, as shown in Figs. 4.2 and 4.3.

3) A graph of the cumulative frequency distribution, representing the empirical cumulative distribution function of r_{XY}, as shown also in Figs. 4.2 and 4.3.

4) The k^{th} moment, m_k, about the mean of the empirical distribution as given by the general formula :

$$m_k = [\sum_{i=1}^{N}(r_i - m_1)^k] / N$$

5) Skewness and kurtosis as given by :

$$\text{Skewness} = m_3 / (m_2)^{3/2}$$

$$\text{Kurtosis} = m_4 / (m_3)^2$$

The simulated distributions of cross-correlation for $\rho_1\rho_2 = 0$, $n = 10, 30$ and $N = 7000$ are shown in Figs. 4.2 and 4.3 , and tabulated in Tables 4.2 and 4.3. Table 4.4 shows the moments, variance, skewness and kurtosis computed for these distributions.

Digressing from the consideration of two processes, consider for a moment the behaviour of an autoregressive process of lag one by itself. Fig. 2.1 shows a series of 100 terms generated according to Eqn.(2.9) with $\rho_1 = .5$ and the corresponding observed autocorrelations $r_{XX}(k)$ of lags $k = 1,2,\ldots,20$. Fig. 2.2 demonstrates these values for $\rho_2 = -.5$. The autocorrelation values of $r_{XX}(k)$ and $r_{YY}(k)$ are given in Table 2.1. The sample autocorrelation of lag k for the X_t process is given by :

$$r_{XX}(k) = \frac{\sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X})}{[\sum_{t=1}^{n-k} (X_t - \bar{X})^2 \sum_{t=1}^{n-k} (X_{t+k} - \bar{X})^2]^{1/2}} \qquad (2.10)$$

The theoretical autocorrelation of lag k is given by

$$\rho_{XX}(k) = \rho^{|k|} \qquad (2.11)$$

For the series with $\rho_1 = .5$, the theoretical autocorrelation function (acf) is given by $\rho_{XX}(k) = .5^{|k|}$, which decays to zero exponentially with increasing lag, as can be observed from the plots of the sample acf in Fig. 2.1(b). For the series with $\rho_2 = -.5$, the theoretical acf is $\rho_{YY}(k) = (-.5)^{|k|}$, which also damps out exponentially but oscillates from positive to negative values, reflecting the oscillatory nature of the series. This behaviour is also indicated by the sample acf in Fig. 2.2. From these two examples it can be seen that a process with only one nonzero autocorrelation, that of lag 1, has an exponentially decaying autocorrelation function. Hence, in practice it is often difficult to determine whether or not the time series comes from an autoregressive model of order one or from some other model. Box and Jenkins [5] have developed guidelines for both identification and estimation of time series models.

Table 2.1

First 20 Lags of the Sample ACF for Sample of 100 Terms of the X_t and Y_t Processes.

k	Sample ACF	
	$r_{XX}(k)$	$r_{YY}(k)$
1	0.4099	-0.3945
2	0.2012	0.1141
3	0.0228	-0.0083
4	0.0069	-0.1193
5	0.0804	-0.0149
6	0.0122	-0.0195
7	0.1349	-0.1372
8	-0.0106	0.0265
9	0.0890	-0.0733
10	-0.0289	0.0403
11	-0.0181	0.1209
12	0.0127	0.0359
13	-0.0428	-0.0617
14	-0.0149	0.0070
15	-0.0447	0.0570
16	0.1209	-0.9440
17	0.4429	0.1043
18	0.0335	-0.1475
19	-0.0895	0.0172
20	-0.0672	-0.0032

Fig.2.1(a)

Sample of 100 terms generated by

$X_t = .5X_{t-1} + Z_t$

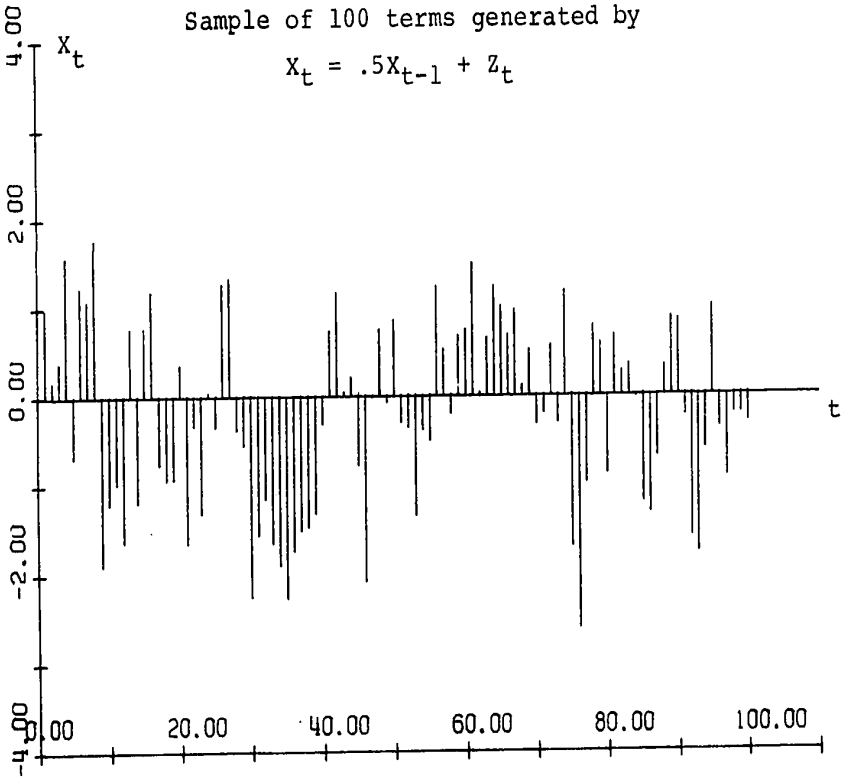


Fig.2.1(b)

First 20 lags of the sample autocorrelation of X_t

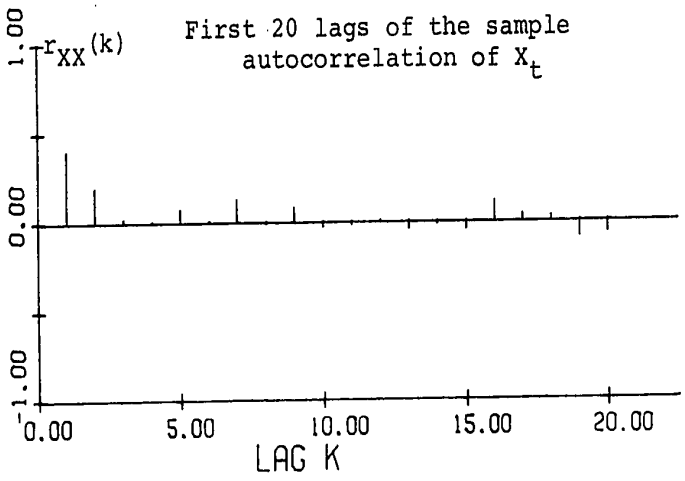


Fig.2.2(a)

Sample of 100 terms generated by

$Y_t = -.5Y_{t-1} + Z_t$

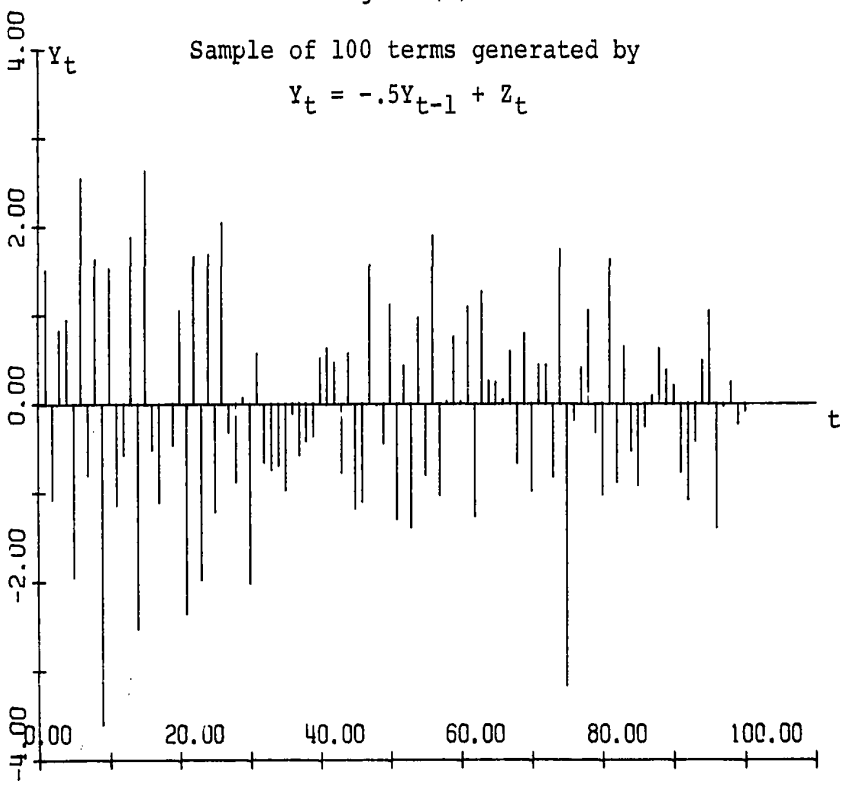
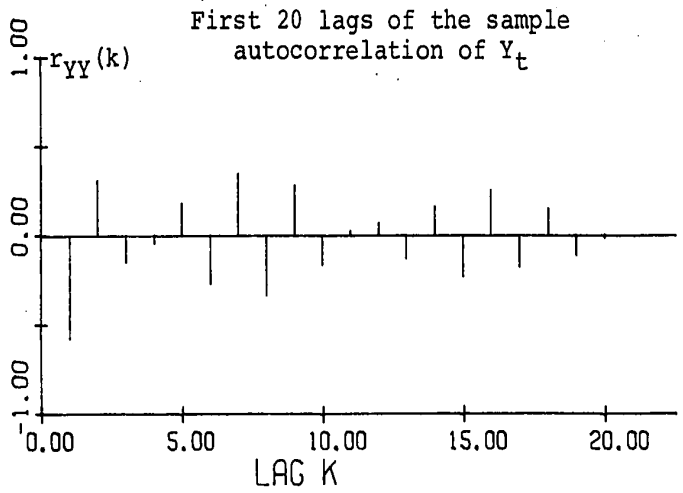


Fig.2.2(b)

First 20 lags of the sample autocorrelation of Y_t



CHAPTER III

PSEUDO-RANDOM NUMBER GENERATION.

3.1 Introduction to the Generation of Random Numbers.

All Monte Carlo methods and most simulation studies depend on the use of random numbers, usually in great quantities. Hence, the need arises for fast methods of generating large number of random numbers with a fairly wide variety of distribution functions.

Most techniques for generating random numbers from specific distributions depend on random numbers uniformly distributed over the interval (0,1). The usual procedure is to generate values of uniform random variables from the interval (0,1) and by some functions transform these values to random numbers from the desired distribution. For example, if y is a uniform random number from the interval (0,1) (that is, U(0,1) random number) then

$$x = - \ln (1 - y) / \lambda$$

is a random number having the exponential distribution with parameter λ. If y_1, y_2 are U(0,1) random numbers then

$$z_1 = (-2 \ln y_1)^{1/2} \cos 2\pi y_2$$

$$z_2 = (-2 \ln y_1)^{1/2} \sin 2\pi y_2 \qquad (3.1)$$

are a pair of independent standard normal (that is, N(0,1)) random numbers. This last set of transfomations is known as the Box-Muller method [6].

In complex sampling experiments it is useful to be able to select random numbers and repeat the calculations as a method of checking the results and increasing the accuracy of the experiment. The need for the generation of a random number sequence that can be regenerated is self-contradictory because of the definition of random numbers. It is, however, possible to obtain random numbers by a deterministic method which display random behaviour. In practice, a random number is obtained usually by a computer program by means of an algorithm which will generate a sequence of numbers satisfying various statistical criteria of randomness. Such a sequence is called pseudo-random.

Two types of methods of generating random numbers for sampling with computers have been proposed, the physical process and the arithmetical process.

In the physical process, the output of some physical device which is attached to the computer, such as random noise or pulse

generated is converted to a sequence of random digits. This method, however, involves practical difficulties such as the storage of a large volume of random numbers for check calculations and hence, is usually not very applicable for fast access of random numbers in present digital machines.

In the arithmetical process, a sequence of pseudo-random numbers is derived using an algorithm and an initial supply of random numbers as starting values. The sequence of generated numbers is always cyclic (since the number representation in a digital computer has a finite number of digits); however, if the cycle is long enough this will present no difficulties. The generated numbers are deterministic and completely predictable as soon as the initial values and computational rules are known. Four types of arithmetical processes for generating pseudo-random numbers have been used in various studies :

a) the mid-square method,
b) the randomization by summation modulo p method,
c) the sequence of digits in transcendental numbers,
d) the residue-class or multiplicative congruential method.

In this thesis only the multiplicative congruential method will be considered (see Section 3.4). Descriptions and discussions on the other three methods may be found in Jannson [17] and Tocher [38].

3.2 Desired Properties of a Pseudo-Random Number Generator.

To determine the performance of a random number generator, a broad range of statistical properties are investigated. Some desirable properties which a 'good' generator should satisfy and which can be used for comparison between different generating processes are :

a) Good statistical behaviour - the generated numbers (in large and small samples) must satisfy various statistical criteria of randomness and the distribution function of these numbers must approximate as closely as possible to the desired distribution.

b) Long period - the period after which the sequence of numbers repeat itself should be sufficiently long to ensure that the sequence contains enough random numbers for it to be useful in a particular problem.

c) Rapid and short calculating procedure - the generating time on the computer must be short and the space used for storage in core memory small.

Points (b) and (c) are often easy to study and for most generators can be achieved by proper choice of initial values

and parameters of the generator. The choice of initial values and parameters, however, will depend on the computer used.

Point (a) is more difficult to study and requires thorough investigation in many different respects. The usual procedure is to test a sample of generated numbers by standard statistical tests concerning distribution and randomness. However, the suitability of a generator for any study depends upon the properties required for that particular use. Hence, a necessary condition for the approval of a sequence of pseudo-random numbers is that the numbers pass such statistical tests as are relevant for the application under consideration.

3.3 Statistical Tests for Pseudo-Random Number Generators.

Ideally these statistical tests should be selected in accordance with the actual applications, since different applications are more or less sensitive to different properties of the random numbers. In practice, however, most of the generators have a general use as standard routines and they consequently have to pass a number of standard tests. The tests most often used will be briefly introduced here. Detailed descriptions of these tests and their applications may be found in [17].

a) Tests for the Distribution of Generated Numbers.

1) Moments.

The mean, variance, third and fourth moments, skewness and kurtosis are computed for samples of the generated numbers and compared with the corresponding values from the true distribution.

2) Goodness-of-Fit of Generated Numbers to a Theoretical Distribution.

A goodness-of-fit test is performed on a sample of the generated numbers to determine how closely the generated numbers fit the desired distribution. The two most commonly used goodness-of-fit tests are :

i) Chi-square test, and

ii) Kolmogorov-Smirnov test.

3) Cumulative Distribution.

The sample cumulative distribution is computed for the generated numbers and compared with the values from the desired theoretical cumulative distribution.

4) Order and Small Sample Statistics.

The r^{th} order-statistic of a sample is the r^{th} smallest observation in a sample. A useful test for the local statistical properties of a generator is to compute the distribution of the order statistics, range, mean and variance for small samples of size, say, 4,6,10,16. A lack of randomness would cause these observed distributions to deviate from the theoretical distributions.

b) Tests for the Randomness of the Generator.

1) Serial Correlation.

A serious type of non-randomness which might be expected from the generator is correlation between successive numbers and between numbers that are k elements apart. Measures of these correlations are the serial correlations of lag 1 and lag k. If the numbers are random there should be no correlation between them. As a test for randomness the sample correlations are computed and compared with the theoretical values.

2) Run Tests.

Serial effects could also be revealed by run tests. A run up (or down) of length p is defined as a subsequence $x_{i-1} > x_i < x_{i+1} < \ldots < x_{i+p-1} < x_{i+p} > x_{i+p+1}$ (and with reversed inequality signs for runs down) in a sequence of n random numbers.

Let r = number of runs in the sequence,

r_p = number of runs of length p in the sequence.

Then expressions for expected values are given by Levene and Wolfowitz [39],

$$E(r) = (2N - 1) / 3 \quad ;$$

$$Var(r) = (16N - 29) / 90 \quad ;$$

$$E(r_p) = \frac{[2N(p^2 + 3p + 1) - 2(p^3 + 3p^2 - p - 4)]}{(p + 3)!}$$

Let $r^{(m)}$ be the number of runs above and below the mean, then

$$E(r^{(m)}) = (N/2) + 1 .$$

These runs are counted by constructing a sequence of N

signs with the i^{th} sign plus or minus depending on whether x is greater than $1/2$ or less than $1/2$. The runs of plus and minus signs are then counted. For testing the generated sequence, the observed moments are compared with the above formulas.

3) Extreme Values.

In a sequence of random numbers the extreme values are expected to be randomly dispersed. The expected number of such values is 1 per 10,000 (see Chen [7]) and they should follow the Poisson distribution with mean 1. In this test the observed cumulative distribution of extreme values is compared with the expected value computed from the Poisson distribution. The Kolmogorov-Smirnov test may be used to compare the two distributions.

3.4 Multiplicative Congruential Pseudo-Random Number Generators.

The multiplicative congruential generator which is based on number theory takes the form

$$x_{n+1} = \lambda x_n \pmod{m}, \quad n=0,1,2,\ldots, \tag{3.2}$$

where λ is the multiplier, m is the modulus, and x_0 is the starting value or seed; x_0, λ and m are selected integers such that $x_0 < m$, $\lambda < m$, $0 < x_0 < m$. The sequence $\{x_n / m\}$ is then taken to be the uniform random number sequence in $(0,1)$.

The procedure used in this generator is briefly outlined as follows :

1) A beginning value x_0 (also known as the seed) is chosen, the first number of the pseudo-random sequence ;

2) x_0 is multiplied by the constant multiplier λ ;

3) The product is divided by the modulus m and the remainder taken as a new x_n , the next pseudo-random number ;

4) Steps 2 and 3 are repeated, with the new random number in place of x_0 , for every successive number desired.

The multiplicative generator is one of the most frequently used generators and appears to be a satisfactory generator for the following reasons :

a) It is economical in computing time and memory requirements. Only two numbers λ and x_n have to be stored. Computation of x_{n+1} can be accomplished by one multiplication since, if m is chosen judiciously, taking moduli merely means shifting the radix point of the product of λ and x_n.

b) For proper choice of m , λ and x_0 , the sequence of random numbers obtained from Eqn. (3.2) will never degenerate (that is, produce any x_n equal to zero, which would make all subsequent x_{n+1} zero), and a maximum cycle length may be obtained.

c) This generator has been well-tested in several studies [7] , [8] , [9] , [26] and has been found to be highly satisfactory - the word 'satisfactory' referring in large part to the existence of good statistical properties for the sequence of numbers generated.

The process of searching for a good multiplicative generator is focused on determining λ and m so that the

generator behaves well according to the evaluation principles of Section 3.2. The following briefly summarizes the conditions which these parameters should satisfy :

1) It can be shown from number theory (Fermat-Euler theorem) that in order to generates a full length cycle (that is, a cycle for which λ^n returns to the starting point for each repetition) :

$$\gcd(\lambda, m) = 1 \quad \text{and} \quad \gcd(x_0, m) = 1$$

2) From number theory it can be shown that if m is a prime number and λ a primitive root of m then Eqn. (3.2) generates a full cycle which is a permutation of the integers $1,2,\ldots,m-1$.

3) The operation of taking a congruence involves a division and it is desirable to choose m to make this usually lengthy operation rapid. For a binary computer the usual choices of m are :

$$m = 2^{\beta} \quad \text{or} \quad m = 2^{\beta} \pm 1$$

where β is the number of bits of a computer word. In these cases, division by m can be replaced by a shift or a shift and a subtraction or addition, respectively.

4) Choice of $m = 2^{\beta}$ makes the calculations fast and simple but it makes it impossible to attain the largest possible

periods. The maximum cycle attainable is of length $2^{\beta-2}$ instead of 2^{β}. The values of λ, x_0 giving this cycle must satisfy the conditions

$$\lambda = \pm 3 \pmod 8$$

$$x_0 \text{ is odd}, \quad 1 \leq x_0 \leq 2^{\beta} - 1 .$$

5) For $m = 2^{\beta} - 1$, the longest cycle (that is, period $m-1$) will be given when β is chosen so that $2^{\beta} - 1$ is a prime and λ is a primitive root of m; then the cycle length is $2^{\beta} - 2$. Note that $2^{\beta} - 1$ is prime only if β is prime.

6) λ should be large so as to prevent a small value x_n being followed by small values λx_n, $\lambda^2 x_n$, $\lambda^3 x_n$, ... A large value of λ causes λx_n to exceed m and so x_{n+1} is not necessarily small.

Some of the multiplicative generators that have been investigated by various authors and tested for use with different types of binary computers are listed below in Table 3.1. In most cases the author has either made reference to or based the choice of the parameters for the generator on the recommendations of Coveyou and MacPherson [8], who have given one of the few analytical evaluations of pseudo-random number generators in the literature.

Table 3.1

Some Pseudo-Random Number Generators for Binary Computers.

Author(s)	Computer	Program Lang.	m	λ	x_0	Gen.* Time (μsec)
S.Gorenstein	n.a.	GPSS	2^{35}	5^{13}	5^{13}	n.a.
D.Y.Downham F.D.K.Roberts	KDF9	ALGOL	67099-547	8192	odd integer	n.a.
G.Marsaglia T.A.Bray	IBM 360 IBM 7094 SRU 1108	FORTRAN FORTRAN FORTRAN	2^{32} 2^{35} 2^{36}	8n±3 8n±3 1	random integer	n.a.
D.W.Hutchinson	IBM 7094	n.a.	$2^{35}-31$	5^5	n.a.	n.a.
P.A.W.Lewis A.S.Goodman J.M.Miller	IBM 360/67	ASSEM.	$2^{31}-1$	7^5	random integer	31.2
D.S.Seraphin	IBM 360/67	ASSEM.	2^{32}	32781	1	11.8
Oakridge Laboratory	IBM 360/67	ASSEM.	2^{32}	45280 -7053	5^{15}	27
E.H.Chen	IBM 360/90	FORTRAN	2^{31}	$2^{14}+3$ $2^{18}+3$	odd integer	35

n.a. stands for 'not available'.

* generating time per number.

In the simulation of the cross-correlation distribution we are required to use standard normal N(0,1) random numbers. Hence, we shall now consider some techniques of generating normal random numbers which are based on the multiplicative congruential method.

3.5 Random Normal Number Generator for Binary Machines.

The random normal number generator which was proposed by E.H.Chen [7] is based on the combination of two multiplicative congruential generators. This dual type of generator has also been suggested by Kronmal [23] and, Maclaren and Marsaglia [28].

The multiplicative congruential generator used by Chen is of the form

$$R_i = R_{i-1}(2^p + K) \pmod{2^{31}} \qquad (3.3)$$

where p is a positive integer, $2 < p < 31$,

K is an odd integer,

R_0 , the starting number, is a random odd integer.

Uniform random variables from (0,1) are then obtained by

$$U_i = R_i \; / \; (2^{31} - 1) \qquad (3.4)$$

Two independent random standard normal variables are produced from two independent uniform variables, U_1 and U_2, by the Box and Muller transformation :

$$X_1 = (-2 \ln U_1)^{1/2} \cos(2\pi U_2)$$
$$X_2 = (-2 \ln U_1)^{1/2} \sin(2\pi U_2) \qquad (3.5)$$

Normal variables with other means and variances may be obtained by suitable linear transformations of the X_i's.

Good statistical behaviour of the generator in Eqn. (3.3) is dependent on the careful choice of the values of p and K. The performance of various combinations of p and K with respect to serial correlation (lag 1), mean and variance for the normal variables was investigated by Chen. In order to improve serial correlation a dual type of generator of the following form was tried and tested by Chen:

$$R_{1,i} = R_{1,i-1}(2^{14} + 3) \pmod{2^{31}}$$
$$R_{2,i} = [R_{2,i-1}(2^{18} + 3)](2^{18} + 3) \pmod{2^{31}} . \qquad (3.6)$$

This generator was found to be satisfactory relative to several criteria for testing normality and randomness. The periods of

the first and second generators are 2^{29} and 2^{28}, respectively. The combined period of the dual generator is at least 2^{31}.

A brief account of the series of statistical tests performed by Chen on his random number generator and the results obtained is given in the following section. To test the generator, ten trials each of sample size $n = 10^6$ were successively generated with the first pair of starting integers randomly chosen. For each trial the series of tests listed below was performed. As a further check on the performance of the generator we repeated some of these tests on the generator using sample sizes of 40,000 , generated with the same pairs of starting integers as those used by Chen. The results obtained were compared with those of Chen's.

A) Tests for Normality

1) Mean and Variance.

The mean and variance were computed for each trial. Assuming the sample mean to be normally distributed with mean zero and variance 1/n, and the sample variance to have a chi-square distribution with n-1 degrees of freedom, Chen found that none of the computed sample means and variances exceeded the .05 level of significance.

We computed the sample mean and variance for 10 trials of sample size 40,000 each and found that the values obtained compared quite well with those obtained by Chen using a sample size of 10^6. The two sets of test results are shown in Table 3.2. Fig. 3.1 illustrates the fit of one of the samples of 40,000 normal random numbers to the standard normal curve.

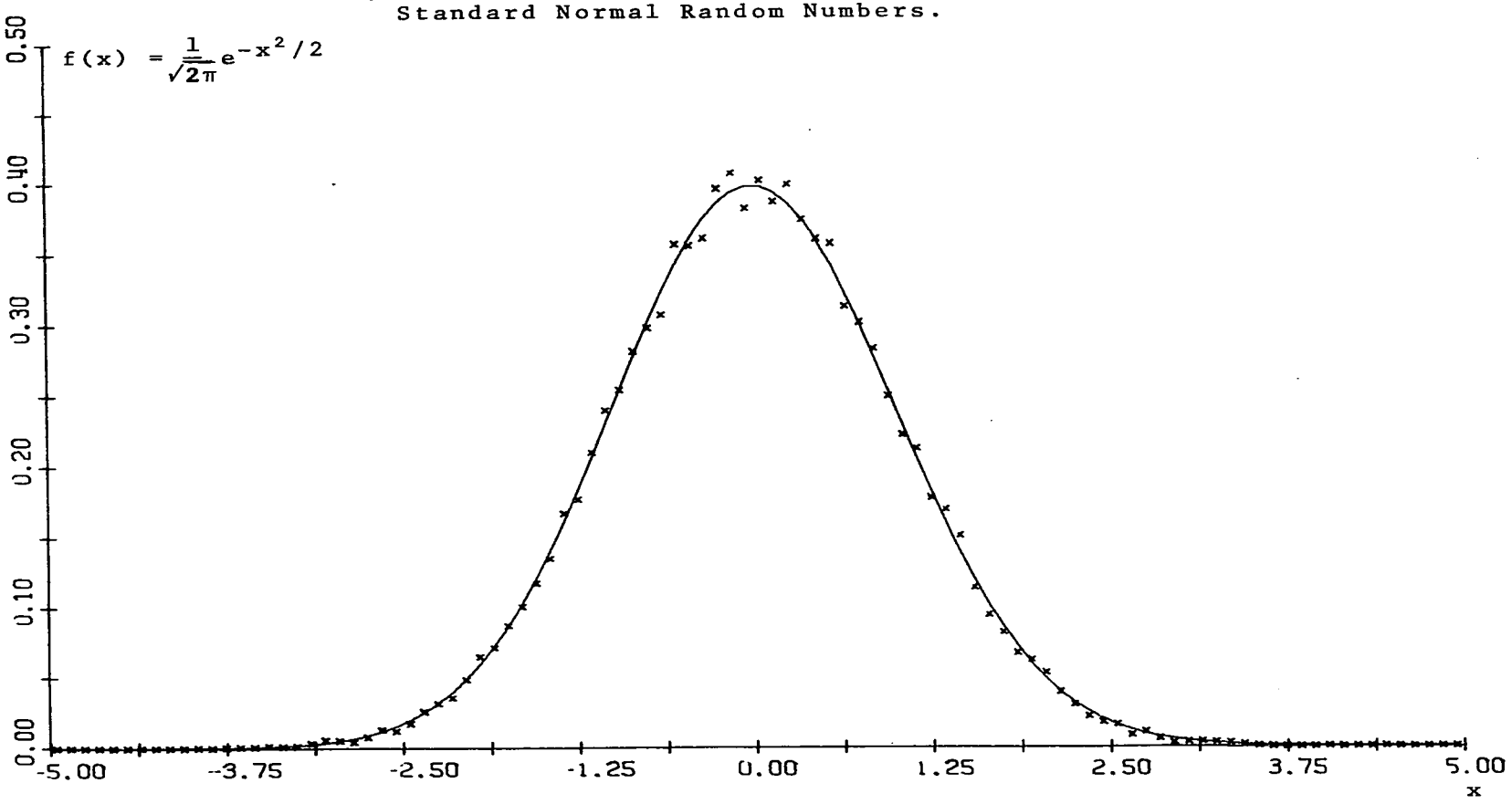
Table 3.2

Mean and Variance for 10 Trials of Sample Size 40,000 and 10^6 each.

Trial	Mean		Variance	
	Sample Size			
	4×10^4	10^6	4×10^4	10^6
1	.0026	-.0001	.9940	.9995
2	-.0028	-.0001	.9976	.9985
3	-.0004	.0005	.9903	1.0010
4	-.0007	.0002	.9928	.9991
5	.0067	.0014	.9969	.9999
6	.0030	.0005	.9985	.9989
7	.0010	-.0010	.9949	1.0014
8	-.0025	-.0003	.9966	1.0014
9	-.0044	-.0006	.9963	.9988
10	-.0080	-.0016	1.0010	1.0020
Mean	-.0006	-.0001	.9959	.9999
Expected Value	0.0	0.0	1.0	1.0

Fig. 3.1

Frequency Distribution of a Sample of 40,000
Standard Normal Random Numbers.



2) Goodness-of-Fit.

The chi-square test was used to test the goodness-of-fit of the generated numbers to the standard normal. The test statistics showed close agreement with the theoretical value and the null hypothesis of normality was accepted at the .05 significance level.

3) Central Area under Normal Curve.

The frequencies of generated deviates between $-x$ and x were compared with the values np, where p is the area under the standard normal curve between $-x$ and x for $x = 0\ (.01)\ 4.50$. There were no large deviations of the observed values from the theoretical. The Kolmogorov statistic was also computed for each trial and was found to lie within the Kolmogorov .10 bound.

4) Order and Small Sample Statistics.

Each trial was divided into $10^6/N$ subsamples for $N = 4,6,10,16$. For each subsample the order statistics, range, mean and variance were computed. The mean and variance of these statistics taken over the $10^6/N$ subsamples showed excellent agreement with the theoretical values.

B) Tests for Randomness

1) Serial Correlation.

Serial correlations of lag 1, 2 and 3 were computed. None of the serial correlations were significantly different from zero at the .05 level.

Using sample sizes of 40,000 we computed the serial correlations of lags 1 and 2 and compared the results with those of Chen's. The two sets of results were found to agree well as can be seen from Table 3.3.

2) Run Tests.

Runs up and down, and runs above and below the mean of various lengths were obtained for each trial. None of the chi-square test statistics computed for these runs were significant at the .05 level.

Table 3.3

Serial Correlations of Lags 1 and 2 for 10 Trials of Sample Size 40,000 and 10^6 each.

Trial	Serial Correlation			
	Lag 1		Lag 2	
	Sample Size			
	4×10^4	10^6	4×10^4	10^6
1	-.0035	-.0001	-.0125	-.0008
2	-.0047	-.0003	.0000	-.0004
3	.0035	-.0001	.0057	-.0004
4	.0002	-.0006	-.0086	-.0005
5	.0033	.0016	.0004	.0015
6	.0016	-.0014	.0006	.0017
7	.0004	.0009	.0039	-.0019
8	-.0006	-.0008	.0062	.0009
9	-.0085	-.0003	-.0086	.0000
10	.0008	-.0006	.0058	-.0006
Mean	-.00075	-.00017	-.0007	-.00005
Expected Value	0.0	0.0	0.0	0.0

3) Extreme Values.

For each trial, 100 subsamples of size 10,000 were examined for values exceeding in absolute value 3.891 (which are classified as extreme values). The observed cumulative distribution of these extreme values agreed excellently with the theoretical distribution.

In a second set of tests, all previously discussed statistics except small sample statistics were computed for each of 100 subsamples of size 10,000. The overall agreement of these test statistics with the theoretical values were quite good. Judging from the results of the statistical tests, the performance of this random number generator may be considered to be remarkably good. We shall next consider the choice of a suitable random number generator for our simulation experiment.

3.6 Choice of a Random Number Generator for Simulation.

The choice of a random number generator for any investigation is based on the properties required for that particular use. In the simulation of the cross-correlation distribution we are interested in simulating realizations of two linear, stationary Markov processes (Eqn. (2.9)) which are uncorrelated. Since correlation between the two processes, X_t and Y_t, could arise only through the terms, Z_t and Z'_t, it is essential that the generator used in the simulation should generate uncorrelated normal randcm numbers. Furthermore, the generated numbers must satisfy in all respects various statistical criteria of normality and randomness. In the process of searching for a suitable generator for the simulation, three recently proposed generators were closely examined - they were

those suggested by the Oak Ridge Laboratory (1968) [32], P.A.W.Lewis (1969) [26] and E.H.Chen (1971) [7]. Some of the important features of these generators are summarized in Table 3.1. Chen's generator has been described in the preceding section. All three generators use the multiplicative congruential method to generate uniform U(0,1) random numbers. Standard normal N(0,1) random numbers are then obtained by the Box-Muller transformation (Eqn. (3.1)).

Of the three generators considered, only the ones proposed by Lewis and Chen have been well-tested for their performances as uniform and normal random number generators, respectively, and have been found to perform remarkably well. Complete descriptions and results of the testing of these generators are available in [7] and [26]. As for the generator suggested by the Oak Ridge Laboratory, there has been no report on its performance or of any testing done on it.

A comparison was made between the three generators using samples of 40,000 standard normal random numbers generated by each of the generators (and the Box-Muller transformation, where necessary). For each sample, the mean, variance, third and fourth moments, skewness and kurtosis were computed, as shown in Table 3.4. Reasonably good results were obtained in each case. For each generator, samples of 1000 generated normal numbers were tested for 'normality' by the Kolmogorov-Smirnov criterion (see Section 4.2 for the Kolmogorov-Smirnov goodness-of-fit

test). In each case there was good agreement between the empirical and theoretical distributions.

Table 3.4

Comparison of Moments, Skewness, Kurtosis of Samples of 40,000 Normal Random Numbers Generated by Various Generators.

Generator	Mean	Var.	3rd Mom.	4th Mom.	Skewness	Kurtosis
Oak.Lab.	.0040	.9975	-.0129	.0140	-.0130	2.999
Lewis	-.0061	.9980	-.0271	.0600	-.0272	2.952
Chen	.0005	.9884	.0047	.0640	.0048	2.995
Expected Value	0	1	0	0	0	3

The above comparison and the information in Table 3.1 did not provide any strong indication as to which of the generators is the more superior or better of the three. It is, however, necessary to decide on one which would be considered most suitable for the simulation. Since the generator proposed by the Oak Ridge Laboratory has not been thoroughly tested and proved satisfactory for application, it was decided to disregard it for the simulation. Between the generators proposed by Lewis and Chen, it was decided after careful consideration that Chen's generator would perhaps be a more suitable choice for the

simulation of the cross-correlation distribution, since

1) the generator generates standard normal random numbers which have been thoroughly tested for normality and randomness using large and small samples, and has been found to be highly satisfactory ;

2) it has been specially designed to minimize serial correlation between generated numbers, which is one important property the generator for the simulation must have ;

3) the period of the generator is at least 2^{31} .

Note that this does not mean that Lewis's generator is in any way inferior to that of Chen's. Using Chen's generator we can avoid having to test the generator thoroughly to ensure it satisfies all properties of normality and randomness, and hence save a considerable amount of computer time. Furthermore, the statistical quality of Chen's generator, as indicated by the test results, are found to be acceptable and sufficient for our simulation purposes. This random number generator is implemented as a Fortran subroutine in the main simulation program. Each call of the subroutine returns a sequence of normal random numbers of specified length.

CHAPTER IV

DISCUSSION OF STATISTICAL TESTS OF SIMULATED DISTRIBUTIONS.

4.1 Introduction to Relevant Statistical Tests of Simulated Distributions.

Before using the simulated distribution for any purpose it is necessary to determine the accuracy of this distribution. The error in a simulation experiment may arise from sampling fluctuations such as variations in sample distributions which are influenced by the sample size, or from nonsampling errors such as nonrandomness of the sample drawn and incorrect population distribution. Hence, the following questions should be considered in the present simulation :

a) How closely does the simulated distribution represent the actual distribution of cross-correlation?

b) How much sampling must be done to reduce the error due to sampling fluctuations to a desired value?

The first question concerns the goodness-of-fit of the simulated distribution while the second question is concerned with the determination of the sample size required to obtain the

simulated distribution to a desired precision.

Since only the distribution of the sample cross-correlation for the case $\rho_1\rho_2 = 0$ is known (it is the null distribution of the Pearson correlation coefficient), it is used as the hypothesized distribution for testing the goodness-of-fit of the simulated distribution. The distribution of the sample cross-correlation with either ρ_1 or ρ_2 equal to zero is simulated and compared with this known distribution. The following tests, which are described in detail in the next section, are used :

1) Goodness-of-fit tests.

a) Kolmogorov-Smirnov test of goodness-of-fit [22,29].

This test compares the empirical cumulative distribution function (cdf) directly with the hypothesized cdf. The measure of discrepancy used is the maximum absolute deviation between these two cdf's. This test statistic can be used to compute the necessary sample size for a desired precision in the simulated distribution, and to form a confidence band which can be used to estimate the accuracy of the approximate distribution given by Eqn. (1.5). This will be explained in the next chapter.

b) Anderson-Darling test of goodness-of-fit [1,25].

Since we are interested in computing the critical points for the sample cross-correlation, it is important that the tails of the simulated distribution agree well with those of the hypothesized distribution. The Anderson-Darling test which is also based on comparing the empirical cdf with the hypothesize cdf is designed to be especially sensitive to discrepancies at the tails of the distributions; and, hence, is used to test for good fitting of the distributions in the tail regions. Details are given in Section 4.3.

2) Comparison of simulated and theoretical critical points.

Another method of checking the accuracy of the simulated distribution is to compare the critical points of this distribution with those of the theoretical distribution. This method is based on Bahadur's [2] theory on sample quantiles and is described in Section 4.4. The theory also furnishes a method for computing error bounds for the critical points of the approximate distribution.

4.2 The Kolmogorov-Smirnov Test for Goodness-of-Fit.

Let X be a random variable with the continuous probability distribution function

$$U(x) = \Pr\{ X < x \} \tag{4.1}$$

Let $X_1, X_2, \ldots, X_N$ be a sample of size N for X, ordered so that $X_1 < X_2 < \ldots < X_N$. The empirical distribution function of the sample $X_1, X_2, \ldots, X_N$ is the step-function $S_N(x)$ defined by

$$S_N(x) = \begin{cases} 0 & \text{for } x < X_1 \\ k/N & \text{for } X_k \leq x < X_{k+1} \\ 1 & \text{for } x > X_N \end{cases} \tag{4.2}$$

That is, $NS_N(x)$ equals the number of variables X_i, which are less than x. For large N one would expect

$$S_N(x) \to U(x) \qquad \text{as } N \to \infty .$$

The goodness-of-fit problem is to devise a test of the hypothesis

$$H_0 : U(x) = F(x) , \quad \text{for all } x ,$$

where $F(x)$ is a completely specified, hypothesized distribution function of the random variable X.

For the purpose of testing the hypothesis H_0, Kolmogorov [22] introduced in 1933 the statistic

$$D(N) = \sup_{-\infty<x<\infty} |S_N(x) - F(x)| , \tag{4.3}$$

which measures the maximum absolute deviation between the sample cumulative distribution $S_N(x)$ and the hypothesized cumulative distribution $F(x)$. H_0 is rejected if $D(N)$ is sufficiently large.

The probability distribution of the random variable $D(N)$ depends on N but is independent of the special form of $F(x)$ provided only that $F(x)$ is continuous (that is, the test is distribution free). The exact distribution of $D(N)$ is not known but Kolmogorov found that $(N^{1/2})D(N)$ has a limiting distribution given by

$$\lim_{N\to\infty} \Pr\{D(N) < zN^{-1/2}\} = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1} e^{-2j^2z^2}$$

$$= L(z) \tag{4.4}$$

The function $L(z)$ has been tabulated by Smirnov [37]. The

early work of Kolmogorov and Smirnov is summarized in [22] and [36]. Tables of the critical values of D(N) have been given by Massey [29], Birnbaum [4] and Miller [31]. Appendix Table D1 gives the critical values of D(N) computed by Massey.

4.3 A Goodness-of-Fit Test for the Tails of a Distribution.

This test of goodness-of-fit was proposed by Anderson and Darling [1] in 1954. The test is sensitive to discrepancies at the tails of the distribution rather than near the median. The test procedure is the following :

Let $x_1 \leq x_2 \leq \ldots \leq x_N$ be N ordered observations in a sample from the random variable X. Let

$$u_i = F(x_i) , \tag{4.7}$$

where F(x) is a completely specified, hypothesized distribution function for X, and let $S_N(x)$ be the empirical distribution function as defined in (4.2). The test criterion suggested by Anderson and Darling is

$$W(N) = N \int_{-\infty}^{\infty} [S_N(x) - F(x)]^2 \, \psi(F(x)) \, dF(x) , \tag{4.8}$$

where $\Psi(F(x)) = \Psi(u)$ is some non-negative weight function chosen to accentuate the values of $S_N(x) - F(x)$ where the test is desired to have sensitivity. The hypothesis that the sample has been drawn from the distribution $F(x)$ is rejected if $W(N)$ is sufficiently large.

For the test to be sensitive to discrepancies at the tails of the distribution $\Psi(u)$ should be large for u near 0 and 1, and small near $u = 1/2$. The weight function chosen by Anderson and Darling is

$$\Psi(u) = 1 / [u(1 - u)] \quad . \tag{4.9}$$

This function has the effect of weighting the tails heavily since it is large near $u = 0$ and $u = 1$.

Subsituting (4.9) for $\Psi(F(x))$, Eqn. (4.8) may be written as

$$\begin{aligned} \frac{1}{N} W(N) &= \int_{-\infty}^{\infty} \frac{[S_N(x) - F(x)]^2}{F(x)[1 - F(x)]} \, dF(x) \\ &= \int_{-\infty}^{x_1} \frac{F^2(x) \, dF(x)}{F(x)[1 - F(x)]} + \int_{x_1}^{x_2} \frac{[S_N(x) - F(x)]^2 \, dF(x)}{S_N(x)[1 - F(x)]} \\ &\quad + \ldots + \int_{x_N}^{\infty} \frac{[1 - F(x)]^2 \, dF(x)}{F(x)[1 - F(x)]} \end{aligned} \tag{4.10}$$

By straightforward integration and collection of terms , (4.10) reduces to

$$W(N) = -N - \frac{1}{N}\sum_{j=1}^{N}(2j-1)[\operatorname{Ln} u_j + \operatorname{Ln}(1-u_{N-j+1})] . \tag{4.11}$$

If this number, W(N), is too large, the hypothesis that F(x) is the true distribution is rejected.

Asymptotic significance (or critical) points for W(N) are given by Anderson and Darling (see Table 4.5). Significance points for W(N) for small sample sizes have been determined and tabulated by Lewis [25] who also gave the following equivalent form of the test statistic W(N) :

$$W(N) = -N - \frac{1}{N}\sum_{j=1}^{N}[(2j-1)\operatorname{Ln} u_j + (2(N-j)+1)\operatorname{Ln}(1-u_j)] \tag{4.12}$$

4.4 Comparison of Simulated and Theoretical Critical Points.

Since the theoretical and approximate density functions of the sample cross-correlation are symmetrical about $r = 0$ in the null case, only the upper (or positive) critical points need to be considered. The critical point of r_{XY} at the α -significance level is also the $(1-\alpha)$-quantile of the cdf of r_{XY} . Bahadur's theory [2] on sample quantiles which is used in the comparison of critical points may be outlined as follows :

Let $F(x)$ be the theoretical probability distribution function. The p-quantile, ξ_p , of $F(x)$ is defined by

$$F(\xi_p) = p , \qquad 0 < p < 1 . \tag{4.13}$$

ξ_p is also the p -lower critical point of $F(x)$ for $0 < p < 1/2$ and the $(1-p)$ -upper critical point of $F(x)$ for $1/2 < p < 1$.

Let

$(X_1 , X_2 , \ldots , X_N)$ be a random sample from $F(x)$,

$Y_{N,p}$ be the sample p-quantile,

B_N be the number of observations X_i in the sample such that $X_i > \xi_p$,

$f(x)$ be the density function corresponding to $F(x)$,

and $q = 1 - p$.

Then,

$$Y_{N,p} = \xi_p + [(B_N - Nq) / Nf(\xi_p)] + R_N \ , \tag{4.14}$$

where R_N becomes neglible as $N \to \infty$. It was shown by Bahadur that with probability one,

$$R_N = O(N^{-3/4} \text{ Log } N) \quad \text{as} \quad N \to \infty \ .$$

Hence, for N large R_N may be assumed to be zero and the term $[B_N - Nq] / Nf(\xi_p)$ gives an estimate of the possible difference between the sample and the theoretical p-quantile.

As a means of checking the accuracy of the simulated distribution, the p-quantile (or q-critical value) can be computed for the simulated and theoretical distributions for $\rho_1\rho_2 = 0$ and the observed difference between the two values can be compared to the corresponding value of the above-stated error estimate which is computed for the two distributions. This test on the simulated distribution will be performed in Section 4.8.

4.5 Error Bounds for Critical Points of Simulated and Approximate Distributions.

The theory on sample quantiles [2] may also be utilized to estimate error bounds for critical points of the simulated distribution and the approximate distribution of Eqn. (1.5). As can be seen from Eqn. (4.14), the error between the theoretical and simulated quantile will not be the same for all quantiles, and, hence, has to be determined individually for each quantile as shown below.

It is known that $Y_{N,p}$ is approximately normally distributed when N is large with

$$E(Y_{N,p}) = \xi_p \quad , \tag{4.15}$$

$$Var(Y_{N,p}) = pq \,/\, Nf^2(\xi_p) \quad ; \tag{4.16}$$

and $N^{1/2}(Y_{N,p} - \xi_p)$ is asymptotically normally distributed when $N \to \infty$ with

$$E(N^{1/2}(Y_{N,p} - \xi_p)) = 0 \quad ; \tag{4.17}$$

$$Var(N^{1/2}(Y_{N,p} - \xi_p)) = pq \,/\, f^2(\xi_p) \quad . \tag{4.18}$$

Let

$$U_{N,p} = Y_{N,p} - \xi_p \quad . \tag{4.19}$$

Since $N^{1/2}U_{N,p}$ is asymptotically normally distributed, we can determine, for various large sample sizes N, certain critical points of the distribution of $U_{N,p}$ as shown below.

Let $Z_{N,p}$ denote the standardized normal variable $N^{1/2}U_{N,p}$. Then

$$\begin{aligned} Z_{N,p} &= [N^{1/2}U_{N,p} - 0] / [\mathrm{Var}(N^{1/2}U_{N,p})]^{1/2} \\ &= [N^{1/2}U_{N,p} f(\xi_p)] / (pq)^{1/2} \quad . \end{aligned} \tag{4.20}$$

Let z_α be the upper α - critical point of the $N(0,1)$ distribution. Then we have,

$$\begin{aligned} P[Z_{N,p} \le z_\alpha] &= 1 - \alpha \quad , \\ P[N^{1/2}U_{N,p} f(\xi_p) / (pq)^{1/2} \le z_\alpha] &= 1 - \alpha \quad , \\ P[N^{1/2}U_{N,p} \le z_\alpha (pq)^{1/2} / f(\xi_p)] &= 1 - \alpha \quad . \end{aligned} \tag{4.21}$$

Hence, the upper α - critical point of the distribution of

$N^{1/2}U_{N,p}$ is given by

$$z_\alpha(pq)^{1/2} / f(\xi_p) \ .$$

It follows that the critical values of the asymptotic distribution of $U_{N,p}$ are given by

$$U_{N,p,\alpha} = z_\alpha(pq)^{1/2} / [N^{1/2} f(\xi_p)] \ . \qquad (4.22)$$

$U_{N,p,\alpha}$ can be used to determine an approximation to the error between the simulated and theoretical critical points of the $p(r)$ distribution. The difference between the simulated and theoretical critical values for a simulation sample size N is compared to the values of $U_{N,p,\alpha}$. If this difference is less than $U_{N,p,\alpha}$ for the specified α, then we are $(1-\alpha)100\%$ certain that the error between the simulated and theoretical critical point is at most $U_{N,p,\alpha}$. Hence, $U_{N,p,\alpha}$ may be taken to be the resultant error in simulation for sample size N.

From Eqn.(4.22) we can compute the sample size which is necessary for a desired precision in the simulated critical points of $p(r)$. For example, to be 99% sure that the error in the simulated .01 - critical point is less than .02, we have,

$$N^{1/2} > z_{.01}[(.99)(.01)]^{1/2} / [(.02)\, f(\xi_{.99})]$$

$$= 2.326[(.99)(.01)]^{1/2} / [.02(.161)]$$

$$= 71.85 .$$

Therefore $N = 5184$.

Hence, for any sample size $N \geq 5200$ used in the simulation the error in simulation for the critical points is at most .02.

4.6 Kolmogorov-Smirnov Test on Simulated Distribution.

To determine the accuracy of the simulation, the distribution of the cross-correlation for the case $\rho_1\rho_2 = 0$ is generated and compared to the known distribution given by Eqn. (1.4) by means of the Kolmogorov-Smirnov test. The following notation is used :

$p_A(r;n,k)$ = approximate distribution p(r) with $\rho_1\rho_2 = k$

$p_S(r;n,k)$ = simulated distribution p(r)

$p_T(r;n,k)$ = theoretical or true distribution p(r)

$F_A(r;n,k)$ = cdf of $p_A(r;n,k)$

$F_S(r;n,k)$ = cdf of $p_S(r;n,k)$

$F_T(r;n,k)$ = cdf of $p_T(r;n,k)$

n = sample size of Markov series

N = number of simulated r values

$D_{ST}(N)$ = $\underset{|r|<1}{\text{Max}} |F_T(r;n,k) - F_S(r;n,k)|$

$D\alpha(N)$ = α -critical value of $D_{ST}(N)$, such that

$$P\{D_{ST}(N) > D_\alpha(N)\} = \alpha ,$$

Using the table of critical values , $D\alpha(N)$, by Massey given in Appendix D (see Table D1.) , we can estimate the sample size necessary for a desired accuracy in the simulated distribution. For a large sample size, N , we can say we are 99% sure that the simulated distribution $F_S(r;n,0)$ will lie within $1.63(N)^{-1/2}$ of the true distribution $F_T(r;n,0)$ over the entire distribution, if the observed maximum deviation $D_{ST}(N)$ is less than the .01 - critical value $D_{.01}(N)$. Therefore to make this statement for a deviation of .02 the necessary sample size is found as follows :

From Appendix Table D1 we find that

$$D_{.01}(N) = 1.63(N)^{-1/2} = .02 ,$$

therefore $N = 6643$.

Hence, for a sample of 7000 the error in the simulated distribution should be within 2% over the entire distribution with 99% probability.

Using sample sizes ranging from 50 to 7000 , $p(r)$ with $\rho_1\rho_2 = 0$ was simulated for $n = 10$, 30 and the Kolmogorov-Smirnov statistic $D_{ST}(N)$ was computed for each of the simulated distributions. The observed values of $D_{ST}(N)$ and corresponding critical values $D_\alpha(N)$ for $\alpha = .01$, $.05$, $.10$ are tabulated in Table 4.1. Fig. 4.1 shows values of $D_{ST}(N)$ plotted against N , for both values of n.

Table 4.1

Kolmogorov - Smirnov Statistics for Simulated Distribution

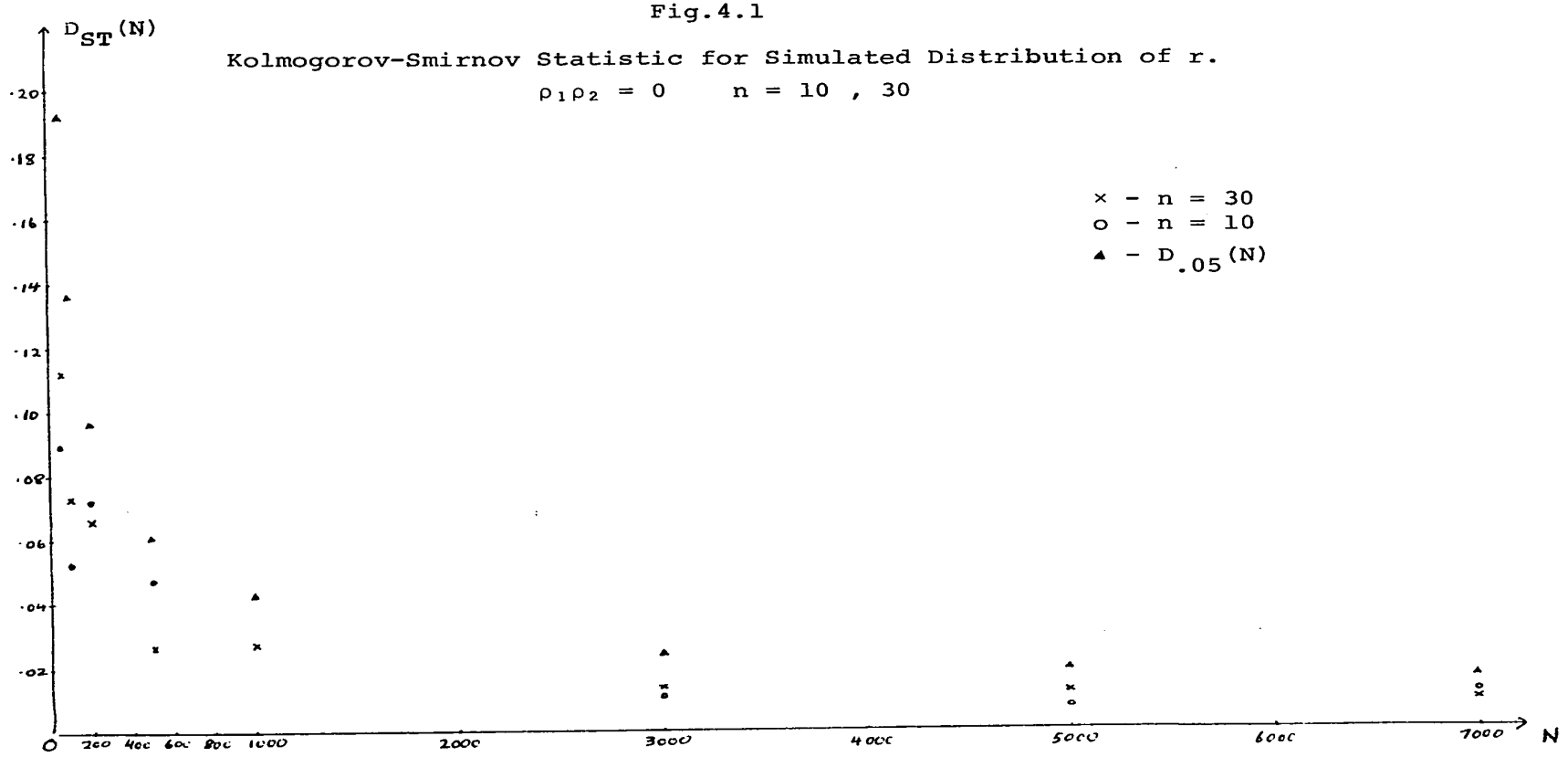
$\rho_1\rho_2 = 0$

Sample Size	Max. Dev. $D_{ST}(N)$		Critical Value $D_\alpha(N)$		
			α		
N	n = 10	n = 30	.01	.05	.10
50	.0894	.1145	.230	.192	.173
100	.0519	.0734	.160	.136	.122
200	.0719	.0660	.115	.096	.087
500	.0478	.0260	.073	.061	.055
1000	.0433	.0272	.051	.043	.039
3000	.0102	.0135	.030	.024	.022
5000	.0072	.0121	.023	.019	.014
7000	.0108	.0094	.020	.016	.015

Fig.4.1

Kolmogorov-Smirnov Statistic for Simulated Distribution of r.

$\rho_1\rho_2 = 0$ n = 10 , 30



As shown in Table 4.1, the observed Kolmogorov-Smirnov statistic, $D_{ST}(N)$, for each of the distributions simulated is less than the corresponding .01-critical value. Using the Kolmogorov-Smirnov criterion, the simulated distribution , $F_S(r;n,0)$, for a sample size of 7000 is within .0108 of the theoretical distribution , $F_T(r;n=10,0)$, and is within .0094 of $F_T(r;n=30,0)$ with 99% probability.

Hence, using sample sizes of 7000 we can estimate with 99% certainty that the error in the simulated distribution will be at most .02 over the entire distribution for any value of $n \geq 6$. (Note that determination of the sample size for this degree of accuracy in the simulated distribution is independent of the value of n.) Having estimated the error in the simulated distribution, we shall now apply this upper bound on the error estimate in the confidence band technique furnished by the Kolmogorov-Smirnov test to evaluate the error in the approximate distribution , $p^*(r)$, for values of $\rho_1\rho_2 \neq 0$ as will be shown in Section 5.1.

Note that the accuracy in the simulated distribution can be improved by using a larger simulation sample size, N. To obtain a maximum of 1% error over the entire simulated distribution would require a sample size of over 25,000. However, this approach is expensive in terms of computer time. Hence, we have limited the value of N used in this simulation to 7000. It is also obvious that it is impossible to obtain very significant

improvement in the accuracy by this direct sampling method. Therefore it is important and profitable to explore (perhaps in further research) alternative methods such as variance - reduction techniques for reducing the sampling variability of simulations and improving the accuracy without increasing the number of simulations.

Tables 4.2 and 4.3 show the frequency distributions of the simulated r values for the case $\rho_1\rho_2 = 0$, for n = 10 and n = 30. Figs. 4.2 and 4.3 illustrate the fit of the simulated values to the theoretical values both for the density function and the cumulative distribution function. Table 4.4 compares the moments, variance, skewness and kurtosis of the simulated distribution to the theoretical distribution. As can be seen from the table, the two sets of values agree quite well.

Note that in order to compare the generated (or simulated) and theoretical p(r) values, the values of the generated p(r) and cdf in columns 4 and 5 of the frequency tables (4.2 and 4.3) are given by ,

Generated p(r) = Frequency / [(N) (Interval Width)]

Generated cdf = Cumulative Sum of Frequency / N

Table 4.2

Simulated and Theoretical Distributions of Cross-Correlation
N = 7000 n = 10 $\rho_1\rho_2 = 0.0$

Range		Freq.	Gen.P(r)	Gen.CDF	Theo.P(r)	Theo.CDF
-1.00	-0.95	2	0.0057	0.0003	0.0001	0.0000
-0.95	-0.90	1	0.0029	0.0004	0.0033	0.0001
-0.90	-0.85	2	0.0057	0.0007	0.0141	0.0005
-0.85	-0.80	13	0.0371	0.0026	0.0356	0.0017
-0.80	-0.75	17	0.0486	0.0050	0.0697	0.0042
-0.75	-0.70	43	0.1229	0.0111	0.1168	0.0088
-0.70	-0.65	56	0.1600	0.0191	0.1765	0.0161
-0.65	-0.60	97	0.2771	0.0330	0.2475	0.0267
-0.60	-0.55	117	0.3343	0.0497	0.3280	0.0410
-0.55	-0.50	127	0.3629	0.0679	0.4157	0.0596
-0.50	-0.45	173	0.4943	0.0926	0.5079	0.0827
-0.45	-0.40	191	0.5457	0.1199	0.6017	0.1104
-0.40	-0.35	249	0.7114	0.1554	0.6942	0.1428
-0.35	-0.30	302	0.8629	0.1986	0.7825	0.1798
-0.30	-0.25	296	0.8457	0.2409	0.8639	0.2210
-0.25	-0.20	324	0.9257	0.2871	0.9359	0.2660
-0.20	-0.15	359	1.0260	0.3384	0.9963	0.3143
-0.15	-0.10	367	1.0490	0.3909	1.0430	0.3654
-0.10	-0.05	364	1.0400	0.4429	1.0750	0.4184
-0.05	0.00	387	1.1060	0.4981	1.0920	0.4727
0.00	0.05	352	1.0060	0.5484	1.0920	0.5273
0.05	0.10	352	1.0060	0.5987	1.0750	0.5816
0.10	0.15	373	1.0660	0.6520	1.0430	0.6346
0.15	0.20	341	0.9743	0.7007	0.9963	0.6857
0.20	0.25	351	1.0030	0.7509	0.9359	0.7340
0.25	0.30	330	0.9429	0.7980	0.8639	0.7791
0.30	0.35	261	0.7457	0.8353	0.7825	0.8202
0.35	0.40	252	0.7200	0.8713	0.6942	0.8572
0.40	0.45	204	0.5829	0.9004	0.6017	0.8896
0.45	0.50	164	0.4686	0.9239	0.5079	0.9173
0.50	0.55	170	0.4857	0.9481	0.4157	0.9404
0.55	0.60	147	0.4200	0.9691	0.3280	0.9590
0.60	0.65	80	0.2286	0.9806	0.2475	0.9733
0.65	0.70	61	0.1743	0.9893	0.1765	0.9839
0.70	0.75	35	0.1000	0.9943	0.1168	0.9912
0.75	0.80	25	0.0714	0.9979	0.0697	0.9958
0.80	0.85	12	0.0343	0.9996	0.0356	0.9984
0.85	0.90	1	0.0029	0.9997	0.0141	0.9995
0.90	0.95	2	0.0057	1.0000	0.0033	0.9999
0.95	1.00	0	0.0000	1.0000	0.0001	1.0000

Table 4.3

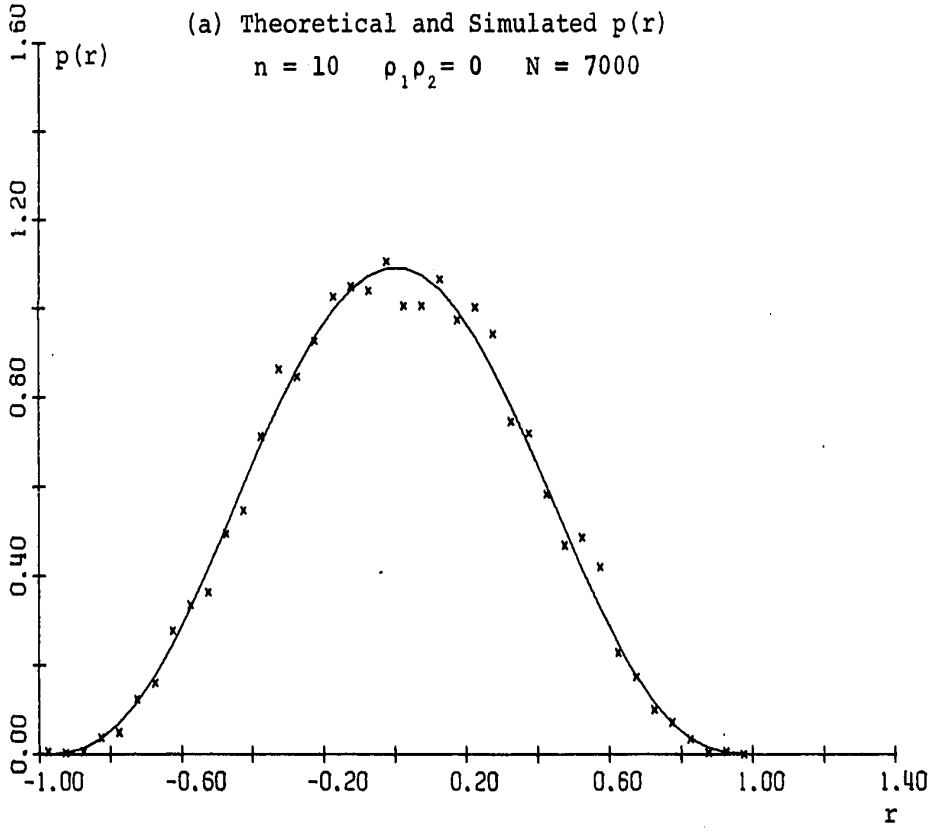
Simulated and Theoretical Distributions of Cross-Correlation
N = 7000 n = 30 $\rho_1\rho_2 = 0.0$

Range		Freq.	Gen.P(r)	Gen.CDF	Theo.P(r)	Theo.CDF
-1.00	-0.95	0	0.0000	0.0000	0.0000	0.0000
-0.95	-0.90	0	0.0000	0.0000	0.0000	0.0000
-0.90	-0.85	0	0.0000	0.0000	0.0000	0.0000
-0.85	-0.80	0	0.0000	0.0000	0.0000	0.0000
-0.80	-0.75	0	0.0000	0.0000	0.0000	0.0000
-0.75	-0.70	0	0.0000	0.0000	0.0001	0.0000
-0.70	-0.65	0	0.0000	0.0000	0.0008	0.0000
-0.65	-0.60	2	0.0057	0.0003	0.0033	0.0001
-0.60	-0.55	6	0.0171	0.0011	0.0113	0.0004
-0.55	-0.50	12	0.0343	0.0029	0.0316	0.0015
-0.50	-0.45	22	0.0629	0.0060	0.0753	0.0040
-0.45	-0.40	52	0.1486	0.0134	0.1570	0.0096
-0.40	-0.35	113	0.3229	0.0296	0.2917	0.0206
-0.35	-0.30	165	0.4714	0.0531	0.4902	0.0399
-0.30	-0.25	246	0.7029	0.0883	0.7527	0.0707
-0.25	-0.20	407	1.1630	0.1464	1.0650	0.1160
-0.20	-0.15	481	1.3740	0.2151	1.3960	0.1775
-0.15	-0.10	577	1.6490	0.2976	1.7050	0.2552
-0.10	-0.05	687	1.9630	0.3957	1.9440	0.3468
-0.05	0.00	720	2.0570	0.4986	2.0750	0.4478
0.00	0.05	684	1.9540	0.5963	2.0750	0.5522
0.05	0.10	691	1.9740	0.6950	1.9440	0.6532
0.10	0.15	603	1.7230	0.7811	1.7050	0.7448
0.15	0.20	494	1.4110	0.8517	1.3960	0.8225
0.20	0.25	389	1.1110	0.9073	1.0650	0.8840
0.25	0.30	252	0.7200	0.9433	0.7527	0.9293
0.30	0.35	179	0.5114	0.9689	0.4902	0.9602
0.35	0.40	100	0.2857	0.9831	0.2917	0.9794
0.40	0.45	63	0.1800	0.9921	0.1570	0.9904
0.45	0.50	38	0.1086	0.9976	0.0753	0.9960
0.50	0.55	10	0.0286	0.9990	0.0316	0.9986
0.55	0.60	3	0.0086	0.9994	0.0113	0.9996
0.60	0.65	2	0.0057	0.9997	0.0033	0.9999
0.65	0.70	1	0.0029	0.9999	0.0008	1.0000
0.70	0.75	1	0.0029	1.0000	0.0001	1.0000
0.75	0.80	0	0.0000	1.0000	0.0000	1.0000
0.80	0.85	0	0.0000	1.0000	0.0000	1.0000
0.85	0.90	0	0.0000	1.0000	0.0000	1.0000
0.90	0.95	0	0.0000	1.0000	0.0000	1.0000
0.95	1.00	0	0.0000	1.0000	0.0000	1.0000

Fig.4.2

(a) Theoretical and Simulated p(r)

n = 10 $\rho_1\rho_2$= 0 N = 7000



(b) Theoretical and Simulated F(r)

n = 10 $\rho_1\rho_2$= 0 N = 7000

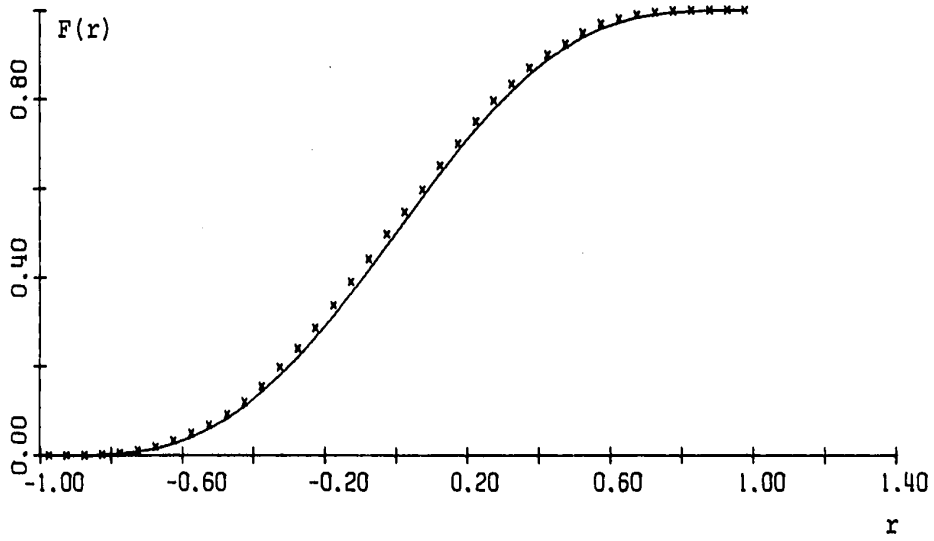
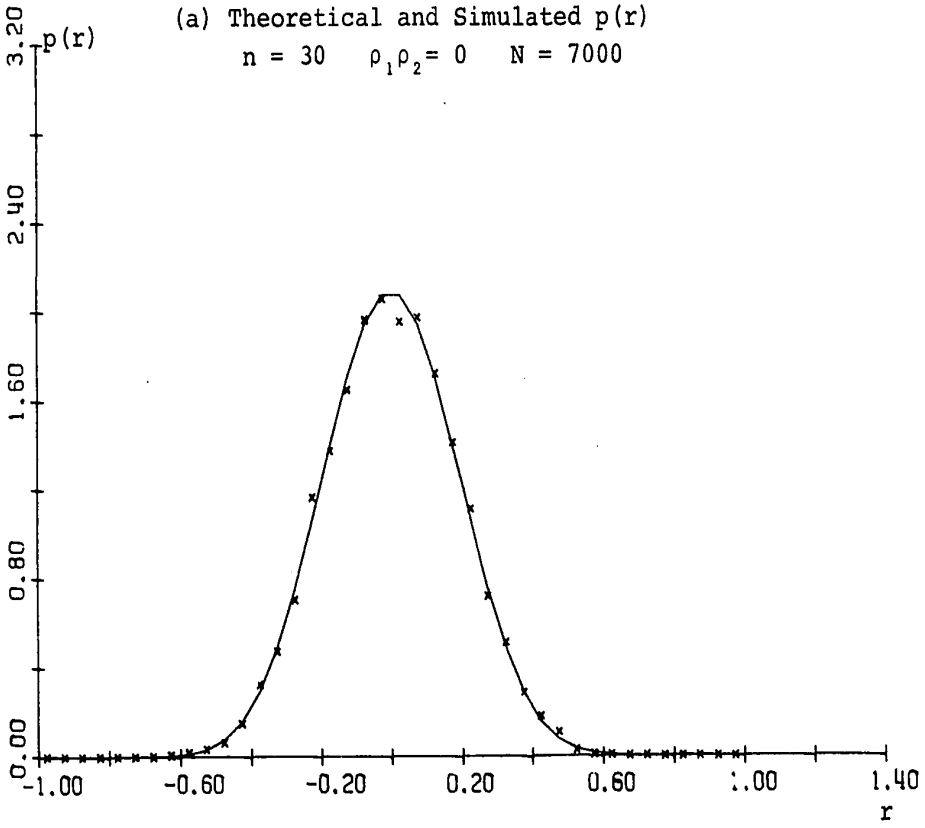


Fig.4.3

(a) Theoretical and Simulated p(r)

n = 30 $\rho_1\rho_2$= 0 N = 7000



(b) Theoretical and Simulated F(r)

n = 30 $\rho_1\rho_2$= 0 N = 7000

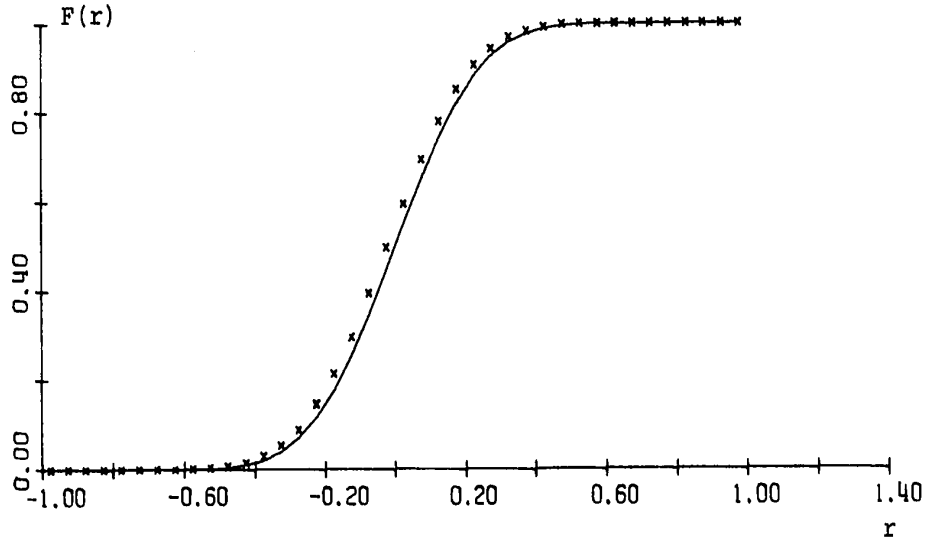


Table 4.4

Moments, Skewness and Kurtosis of Simulated and Theoretical Distribution.

$\rho_1\rho_2 = 0$ N = 7000

	n = 10		n = 30	
	Theo.	Sim.	Theo.	Sim.
Mean	0.0000	0.0042	0.0000	0.0018
Variance	0.1111	0.1108	0.0345	0.0350
Third Moment	0.0000	-.0050	0.0000	0.0001
Fourth Moment	0.0303	0.0297	0.0033	0.0035
Skewness	0.0000	-.0136	0.0000	0.0202
Kurtosis	2.4550	2.4215	2.8070	2.8259

4.7 Anderson-Darling Test on Simulated Distribution.

One of our main objectives is to determine the critical values of the distribution of cross-correlation. We are therefore interested in the tails of our simulated distribution - those areas at the ends of the range containing 1% to 5% of the total area under the curve. To obtain good estimates of the critical values, the simulated distribution should have a good fit at the tail regions. Since the Kolmogorov-Smirnov test is not particularly sensitive to discrepancies at the tails of the distribution, the Anderson-Darling test was used to check the goodness-of-fit at these regions of the simulated distribution. The following notation is used :

N = number of simulations of r values

r_i = i^{th} observation in a sample of N simulations

u_i = $F_T(r_i;n,k)$ (theoretical cdf)

$W(N)$ = Anderson-Darling statistic given by

$$W(N) = -N - \frac{1}{N}\sum_{j=1}^{N}(2j-1)[\ln u_j + \ln(1-u_{N-j+1})]$$

$W_\alpha(N)$ = α - critical value of $W(N)$

The above test was performed on the simulated distributions with $\rho_1\rho_2 = 0$ for $n = 10, 30$, using the known theoretical distribution of Eqn.(1.4). Results of the test appear in Table 4.5.

Table 4.5

Anderson-Darling Statistic for Simulated Distribution

$\rho_1\rho_2 = 0$

	$W(N)$		$W_\alpha(N)$	
N	n = 10	n = 30	α = .01	α = .05
5	1.542	.971	3.95	2.53
8	2.408	.879	3.95	2.52
10	2.139	1.112	3.85	2.49
100	2.152	.646	3.85	2.49

The values of W(N) obtained for both values of n are well below the the 5% significance point for all cases of N. Hence, we may conclude that the simulated distribution has a good fit at the tail regions for all values of $n \geq 6$.

4.8 Comparison of Simulated and Theoretical Critical Points.

Using Bahadur's theory [2] on sample quantiles described in Section 4.5 , a comparison is made between the critical points of the simulated and theoretical distributions for $\rho_1\rho_2 = 0$, as a final check on the accuracy of the simulation. The following notation (with reference to Section 4.5) is used :

$q = 1 - p$

$r_{T,q}(n,k)$ = true q - critical value of r for $\rho_1\rho_2 = k$, that is,

$$Pr[r \leq r_{T,q}(n,k)] = 1 - q = p$$

$r_{A,q}(n,k)$ = approximate q - critical value of r

$r_{S,q}(n,k)$ = simulated q - critical value of r

N = number of simulations of r

B_N = number of observations in a sample of r that are greater than $r_{T,q}(n,k)$

z_α = α - critical value of the $N(0,1)$ distribution, that is,

$$\Pr[z \leq z_\alpha] = 1 - \alpha$$

$$u_{N,q} = | r_{S,q}(n,k) - r_{T,q}(n,k) |$$

$u_{N,q,\alpha}$ = α - critical value of $u_{N,q}$, that is,

$$\Pr[u_{N,q} \leq u_{N,q,\alpha}] = 1 - \alpha$$

Rewriting Eqn.(4.14) in terms of the q - critical value of r, we have,

$$r_{S,q}(n,k) = r_{T,q}(n,k) + [(B_N - Nq) / Np_T(r_{T,q};n,k)] + R_N, \tag{4.22}$$

where with probability one,

$$R_N = O(N^{-3/4} \text{ Log } N)$$

becomes neglible as $N \to \infty$.

Assuming R_N to be zero for large N, the term

$$| [B_N - Nq] / Np_T(r_{T,q};n,k) |$$

gives an estimate of the possible difference between the simulated and theoretical q - critical values. This term may be compared with the observed difference,

$$u_{N,q} = | r_{S,q}(n,k) - r_{T,q}(n,k) |,$$

as a final check on the performance of the simulation. The theoretical and observed differences for the case $\rho_1\rho_2 = 0$, $n = 10, 30$, $N = 7000$ and $q = .01, .02, .05$ are shown in Table 4.6.

Table 4.6

Theoretical and Observed Difference between Simulated and Theoretical Critical Values

$\rho_1\rho_2 = 0$, $N = 7000$

Series Sample Size n	Sig. - Level q	Sim. Crit. Value $r_{S,q}$	Theo. Crit. Value $r_{T,q}$	Obser. Diff. $u_{N,q} = \lvert r_{T,q} - r_{S,q} \rvert$	Theo. Diff. $\left\lvert \frac{B_N - Nq}{N p_T (r_{T,q})} \right\rvert$
10	.01	.7094	.7154	.0061	.0056
	.02	.6493	.6546	.0053	.0028
	.05	.5496	.5494	.0002	.0049
30	.01	.4310	.4226	.0084	.0123
	.02	.3798	.3770	.0028	.0085
	.05	.3081	.3061	.0020	.0113

Table 4.6 shows that the observed differences are less than the theoretical differences in most cases and do not differ too greatly from the theoretical values in others. This serves to indicate reasonably good accuracy in the simulated distribution.

4.9 Error Bounds for Simulated Critical Values.

An error bound for the simulated q - critical value for a particular set of parameters $(n, \rho_1\rho_2)$ can be determined using the theory described in Section 4.5. It was shown that the α - critical value , $u_{N,q,\alpha}$, of the difference between the theoretical and simulated critical value, $u_{N,q}$, is given by

$$u_{N,q,\alpha} = z_\alpha (pq)^{1/2} / [N^{1/2} p_T(r_{T,q}; n, \rho_1\rho_2)] \quad , \qquad (4.23)$$

where $q = 1 - p$ and z_α is the α - critical value of the N(0,1) distribution (as defined in the previous section). The values of $u_{N,q,\alpha}$ for $\rho_1\rho_2 = 0$, $\alpha = .01$, $.02$, $.05$, $q = .01$, $.02$, $.05$, $n = 10$, 30 and $N = 7000$ are shown in Table 4.7.

Comparison of the observed differences , $u_{N,q}$, in column 5 of Table 4.6 with the critical values of Table 4.7 shows that in all cases , the observed values of $u_{N,q}$ are much less than the $u_{N,q,\alpha}$ values. Hence, we can, for example, say with 99% probability that the error in the simulated .01 - critical value for $n = 7000$, $n = 30$ is not more than $u_{N,q,.01} = .0172$. That is , we may conclude that

$$r_{T,.01} = r_{S,.01} \pm .0172$$

Table 4.7

Critical Points of $u_{N,q} = | r_{T,q} - r_{S,q} |$

$\rho_1\rho_2 = 0$, $N = 7000$

Series Sample Size	Sig. - Level	Theo. Crit. Value	$u_{N,q,\alpha}$ Significance Level α		
n	q	$r_{T,q}$	.01	.02	.05
10	.01	.7154	.0218	.0192	.0154
	.02	.6546	.0191	.0168	.0135
	.05	.5494	.0163	.0144	.0115
30	.01	.4226	.0172	.0152	.0121
	.02	.3770	.0136	.0120	.0096
	.05	.3061	.0105	.0092	.0074

CHAPTER V

DETERMINATION OF THE ACCURACY OF THE APPROXIMATE DISTRIBUTION AND ITS CRITICAL VALUES.

5.1 Confidence Band for Distribution Functions.

One of the most useful features of the Kolmogorov-Smirnov test is that the critical values of the test statistic, D(N) (Eqn. (4.3)) may be used to set a confidence band for an unknown continuous distribution function.

Let $F(x)$ be the true unknown distribution function,

$S_N(x)$ be the sample distribution function based on a sample drawn from $F(x)$,

$D_\alpha(N)$ be the α-critical value of $D(N)$, that is,

$$\Pr[\, D(N) \geq D_\alpha(N) \,] = \alpha \quad .$$

Then,

$$\Pr[\, D(N) = \sup_x |S_N(x) - F(x)| > D_\alpha(N) \,] = \alpha \quad . \tag{5.1}$$

That is,

$$\Pr[\, S_N(x) - D_\alpha(N) \leq F(x) \leq S_N(x) + D_\alpha(N) \ ; \text{for all } x] = 1 - \alpha \ . \tag{5.2}$$

Thus for any unknown distribution function $F(x)$, a random sample could be drawn from the distribution and a confidence band of width $\pm D_\alpha(N)$ set up around the sample distribution function $S_N(x)$, so that with probability $1-\alpha$ the true $F(x)$ lies entirely within this band. This is a simple and direct method of estimating a distribution function.

We shall now apply this method to compare the approximate and simulated distributions for a value of $\rho_1\rho_2 \neq 0$, and estimate the error in the approximate distribution. The procedure is described in the following section.

5.2 Determination of Accuracy of Approximate Distribution.

By simulating the distribution of the cross-correlation, r_{XY}, for the case $\rho_1\rho_2 = 0$ and comparing it to the known theoretical distribution of Eqn. (1.4) using the Kolmogorov-Smirnov criterion, we estimated the error in simulation to be at most .02 over the entire distribution. Since this error will mainly be due to sampling fluctuations (that is, variations in simulation results which are dependent on the size and number of samples of r taken) we shall assume that the same error bound holds in the simulation of $p(r)$ for non-zero values of $\rho_1\rho_2$. That is, the error in the simulated distribution with $\rho_1\rho_2 = k \neq 0$, $-1 < k < 1$, using a simulation sample of 7000 r

values is within ±.02 of the true distribution of r for $\rho_1\rho_2 = k$ with probability .99. This assumption can be justified by showing that the assumption holds true for some known distribution of a form similar to p(r) and simulated by means of the same random number generator. The justification of this assumption by showing that it holds true for the normal distribution is given in Appendix B. We now proceed to make an estimate of the accuracy of the approximate distribution, $p^*(r)$, given by Eqn. (1.5), using the confidence band technique described in the previous section.

Having estimated the error in simulation to be at most .02 for sample sizes of 7000, we now simulate the distribution of r for $\rho_1\rho_2 = k \neq 0$, $-1 < k < 1$, and a given value of n, using this sample size. We then set up a confidence band of width ±.02 around the simulated cumulative distribution $F_S(r;n,k)$ (refer Section 4.6 for notation). We can say with 99% certainty that the true cdf $F_T(r;n,k)$ will lie within this confidence band, that is, within ±.02 of the simulated distribution $F_S(r;n,k)$. The resulting probability statement, for the given values of n and k, and all values of r, is

$$\Pr[F_S(r;n,k) - .02 < F_T(r;n,k) < F_S(r;n,k) + .02] = .99 . \tag{5.3}$$

Next, we evaluate the approximate cdf, $F_A(r;n,k)$, using the approximate density function, $p^*(r)$, of Eqn. (1.5) with $\rho_1\rho_2 = k$, and compute the maximum deviation between $F_A(r;n,k)$ and $F_S(r;n,k)$, which is given by

$$D_{SA}(N;n,k) = \max_{|r|<1} |F_A(r;n,k) - F_S(r;n,k)| \quad . \tag{5.4}$$

If $D_{SA}(N;n,k) < .02$, $F_A(r;n,k)$ lies within the .02 confidence band of $F_S(r;n,k)$. The deviation between the approximate and true distribution can then be at most $\pm.04$, that is

$$|F_A(r;n,k) - F_T(r;n,k)| < .04 .$$

Thus we can conclude with 99% certainty that the error in the approximate distribution is less than .04.

If $D_{SA}(N;n,k) > .02$, $F_A(r;n,k)$ falls outside the confidence band and we may say that the error in the approximate distribution could be greater than .04.

The above test was performed on the approximate distribution for the following values of the parameters :

$\rho_1 = .2$	$\rho_2 = .5$	$\rho_1\rho_2 = .10$	$n = 10, 30$
$\rho_1 = -.2$	$\rho_2 = .5$	$\rho_1\rho_2 = -.10$	$n = 30$
$\rho_1 = -.7$	$\rho_2 = .7$	$\rho_1\rho_2 = -.49$	$n = 30$

$\rho_1 = -.8 \quad \rho_2 = .9 \quad \rho_1\rho_2 = -.72 \quad n = 30$

$\rho_1 = -.9 \quad \rho_2 = .9 \quad \rho_1\rho_2 = -.81 \quad n = 30$

$\rho_1 = .7 \quad \rho_2 = .7 \quad \rho_1\rho_2 = .49 \quad n = 30$

$\rho_1 = .8 \quad \rho_2 = .9 \quad \rho_1\rho_2 = .72 \quad n = 30$

$\rho_1 = .9 \quad \rho_2 = .9 \quad \rho_1\rho_2 = .81 \quad n = 30$

The results of the test are tabulated in Table 5.1. Tables 5.2 - 5.6 show the frequency distributions of the simulated r values for the parameters $(n, \rho_1\rho_2) = (10,.1), (30,.1), (30,.49), (30,-.1), (30,-.49)$. Figs. 5.1 - 5.9 illustrate the fit between the simulated and approximate distributions, for all the parameters considered. Tables 5.7 - 5.9 compare the moments, skewness and kurtosis of the approximate distributions with the corresponding values of the simulated distributions.

Table 5.1

Maximum Deviation between Approximate and Simulated Distributions

N = 7000

n	ρ_1	ρ_2	$\rho_1\rho_2$	$D_{SA}(N)$
10	.2	.5	.10	.0102
30	.2	.5	.10	.0099
30	.7	.7	.49	.0152
30	.8	.9	.72	.0322
30	.9	.9	.81	.0395
30	-.2	.5	-.10	.0078
30	-.7	.7	-.49	.0120
30	-.8	.9	-.72	.0571
30	-.9	.9	-.81	.1744

Table 5.2

Simulated and Approximate Distributions of Cross-Correlation
N = 7000 n = 10 $\rho_1\rho_2$ = 0.1

Range		Freq.	Gen.P(r)	Gen.CDF	Appr.P(r)	Appr.CDF
-1.00	-0.95	0	0.0000	0.0000	0.0001	0.0000
-0.95	-0.90	1	0.0029	0.0001	0.0042	0.0001
-0.90	-0.85	8	0.0229	0.0013	0.0184	0.0006
-0.85	-0.80	13	0.0371	0.0031	0.0464	0.0021
-0.80	-0.75	38	0.1086	0.0086	0.0888	0.0055
-0.75	-0.70	39	0.1114	0.0141	0.1446	0.0112
-0.70	-0.65	93	0.2657	0.0274	0.2219	0.0201
-0.65	-0.60	98	0.2800	0.0414	0.2879	0.0326
-0.60	-0.55	114	0.3257	0.0577	0.3700	0.0490
-0.55	-0.50	148	0.4229	0.0789	0.4553	0.0696
-0.50	-0.45	186	0.5314	0.1054	0.5413	0.0946
-0.45	-0.40	228	0.6514	0.1380	0.6253	0.1237
-0.40	-0.35	250	0.7143	0.1737	0.7053	0.1570
-0.35	-0.30	257	0.7343	0.2104	0.7794	0.1942
-0.30	-0.25	275	0.7857	0.2497	0.8458	0.2348
-0.25	-0.20	321	0.9171	0.2956	0.9032	0.2786
-0.20	-0.15	316	1.9029	0.3407	0.9505	0.3250
-0.15	-0.10	354	1.0110	0.3913	0.9867	0.3735
-0.10	-0.05	364	1.0400	0.4433	1.0110	0.4235
-0.05	0.00	362	1.0340	0.4950	1.0240	0.4744
0.00	0.05	348	0.9943	0.5447	1.0240	0.5256
0.05	0.10	329	0.9400	0.5917	1.0110	0.5765
0.10	0.15	374	1.0690	0.6451	0.9867	0.6265
0.15	0.20	344	0.9829	0.6943	0.9505	0.6750
0.20	0.25	300	0.8571	0.7371	0.9032	0.7214
0.25	0.30	304	0.8686	0.7806	0.8458	0.7652
0.30	0.35	301	0.8600	0.8236	0.7794	0.8058
0.35	0.40	260	0.7429	0.8607	0.7053	0.8430
0.40	0.45	219	0.6257	0.8920	0.6253	0.8763
0.45	0.50	164	0.4686	0.9154	0.5413	0.9054
0.50	0.55	166	0.4743	0.9391	0.4553	0.9304
0.55	0.60	128	0.3657	0.9574	0.4553	0.9304
0.60	0.65	89	0.2543	0.9701	0.3700	0.9510
0.65	0.70	82	0.2343	0.9819	0.2879	0.9674
0.70	0.75	63	0.1800	0.9909	0.1446	0.9888
0.75	0.80	38	0.1086	0.9963	0.0888	0.9945
0.80	0.85	12	0.0343	0.9980	0.0463	0.9979
0.85	0.90	11	0.0314	0.9996	0.0184	0.9994
0.90	0.95	3	0.0086	1.0000	0.0042	0.9999
0.95	1.00	0	0.0000	1.0000	0.0001	1.0000

Table 5.3

Simulated and Approximate Distributions of Cross-Correlation
N = 7000 n = 30 $\rho_1\rho_2$ = 0.1

Range		Freq.	Gen.P(r)	Gen.CDF	Appr.P(r)	Appr.CDF
-1.00	-0.95	0	0.0000	0.0000	0.0000	0.0000
-0.95	-0.90	0	0.0000	0.0000	0.0000	0.0000
-0.90	-0.85	0	0.0000	0.0000	0.0000	0.0000
-0.85	-0.80	0	0.0000	0.0000	0.0000	0.0000
-0.80	-0.75	0	0.0000	0.0000	0.0001	0.0000
-0.75	-0.70	0	0.0000	0.0000	0.0004	0.0000
-0.70	-0.65	1	0.0029	0.0001	0.0022	0.0001
-0.65	-0.60	2	0.0057	0.0004	0.0080	0.0003
-0.60	-0.55	13	0.0371	0.0023	0.0234	0.0010
-0.55	-0.50	19	0.0543	0.0050	0.0568	0.0029
-0.50	-0.45	52	0.1486	0.0124	0.1194	0.0072
-0.45	-0.40	93	0.2657	0.0257	0.2226	0.0156
-0.40	-0.35	131	0.3743	0.0444	0.3749	0.0303
-0.35	-0.30	194	0.5543	0.0721	0.5782	0.0539
-0.30	-0.25	277	0.7914	0.1117	0.8254	0.0888
-0.25	-0.20	400	1.1430	0.1689	1.0990	0.1369
-0.20	-0.15	476	1.3600	0.2369	1.3730	0.1987
-0.15	-0.10	601	1.7170	0.3227	1.6170	0.2737
-0.10	-0.05	624	1.7830	0.4119	1.8010	0.3594
-0.05	0.00	652	1.8630	0.5050	1.8990	0.4523
0.00	0.05	662	1.8910	0.5996	1.8990	0.5477
0.05	0.10	604	1.7260	0.6859	1.8010	0.6406
0.10	0.15	569	1.6260	0.7671	1.6170	0.7263
0.15	0.20	466	1.3310	0.8337	1.3730	0.8013
0.20	0.25	396	1.1310	0.8903	1.0990	0.8631
0.25	0.30	297	0.8486	0.9327	0.8254	0.9112
0.30	0.35	202	0.5771	0.9616	0.5782	0.9461
0.35	0.40	115	0.3286	0.9780	0.3749	0.9697
0.40	0.45	87	0.2486	0.9904	0.2226	0.9845
0.45	0.50	37	0.1057	0.9957	0.1194	0.9928
0.50	0.55	13	0.0371	0.9976	0.0568	0.9971
0.55	0.60	13	0.0371	0.9994	0.0234	0.9990
0.60	0.65	3	0.0086	0.9999	0.0080	0.9997
0.65	0.70	0	0.0000	0.9999	0.0022	0.9999
0.70	0.75	0	0.0000	0.9999	0.0004	1.0000
0.75	0.80	1	0.0029	1.0000	0.0000	1.0000
0.80	0.85	0	0.0000	1.0000	0.0000	1.0000
0.85	0.90	0	0.0000	1.0000	0.0000	1.0000
0.90	0.95	0	0.0000	1.0000	0.0000	1.0000
0.95	1.00	0	0.0000	1.0000	0.0000	1.0000

Table 5.4

Simulated and Approximate Distributions of Cross-Correlation
N = 7000 n = 30 $\rho_1\rho_2$ = 0.49

Range		Freq.	Gen.P(r)	Gen.CDF	Appr.P(r)	Appr.CDF
-1.00	-0.95	0	0.0000	0.0000	0.0000	0.0000
-0.95	-0.90	0	0.0000	0.0000	0.0000	0.0000
-0.90	-0.85	0	0.0000	0.0000	0.0002	0.0000
-0.85	-0.80	1	0.0029	0.0001	0.0025	0.0001
-0.80	-0.75	1	0.0029	0.0003	0.0112	0.0004
-0.75	-0.70	7	0.0200	0.0013	0.0323	0.0014
-0.70	-0.65	23	0.0657	0.0045	0.0708	0.0039
-0.65	-0.60	37	0.1057	0.0099	0.1299	0.0088
-0.60	-0.55	56	0.1600	0.0178	0.2099	0.0172
-0.55	-0.50	107	0.3057	0.0331	0.3088	0.0301
-0.50	-0.45	135	0.3857	0.0524	0.4227	0.0484
-0.45	-0.40	186	0.5314	0.0790	0.5466	0.0726
-0.40	-0.35	244	0.6971	0.1139	0.6750	0.1031
-0.35	-0.30	273	0.7800	0.1529	0.8021	0.1400
-0.30	-0.25	296	0.8457	0.1951	0.9224	0.1832
-0.25	-0.20	353	1.0086	0.2456	1.0308	0.2321
-0.20	-0.15	445	1.2714	0.3091	1.1231	0.2860
-0.15	-0.10	405	1.1571	0.3670	1.1955	0.3440
-0.10	-0.05	441	1.2600	0.4300	1.2454	0.4052
-0.05	0.00	445	1.2714	0.4936	1.2708	0.4682
0.00	0.05	467	1.3343	0.5603	1.2708	0.5318
0.05	0.10	438	1.2514	0.6229	1.2454	0.5948
0.10	0.15	448	1.2800	0.6869	1.1955	0.6559
0.15	0.20	402	1.1486	0.7443	1.1231	0.7140
0.20	0.25	370	1.0571	0.7971	1.0308	0.7679
0.25	0.30	345	0.9857	0.8464	0.9224	0.8168
0.30	0.35	283	0.8086	0.8869	0.8021	0.8599
0.35	0.40	207	0.5914	0.9164	0.6750	0.8969
0.40	0.45	190	0.5429	0.9436	0.5466	0.9274
0.45	0.50	123	0.3514	0.9611	0.4227	0.9516
0.50	0.55	116	0.3314	0.9777	0.3088	0.9699
0.55	0.60	81	0.2314	0.9893	0.2099	0.9828
0.60	0.65	43	0.1229	0.9954	0.1299	0.9912
0.65	0.70	21	0.0600	0.9984	0.0708	0.9961
0.70	0.75	8	0.0229	0.9996	0.0323	0.9986
0.75	0.80	3	0.0086	1.0000	0.0112	0.9996
0.80	0.85	0	0.0000	1.0000	0.0025	0.9999
0.85	0.90	0	0.0000	1.0000	0.0002	1.0000
0.90	0.95	0	0.0000	1.0000	0.0000	1.0000
0.95	1.00	0	0.0000	1.0000	0.0000	1.0000

Table 5.5

Simulated and Approximate Distributions of Cross-Correlation
N = 7000 n = 30 $\rho_1\rho_2 = -.49$

Range		Freq.	Gen.P(r)	Gen.CDF	Appr.P(r)	Appr.CDF
-1.00	-0.95	0	0.0000	0.0000	0.0000	0.0000
-0.95	-0.90	0	0.0000	0.0000	0.0000	0.0000
-0.90	-0.85	0	0.0000	0.0000	0.0000	0.0000
-0.85	-0.80	0	0.0000	0.0000	0.0000	0.0000
-0.80	-0.75	0	0.0000	0.0000	0.0000	0.0000
-0.75	-0.70	0	0.0000	0.0000	0.0000	0.0000
-0.70	-0.65	0	0.0000	0.0000	0.0000	0.0000
-0.65	-0.60	0	0.0000	0.0000	0.0000	0.0000
-0.60	-0.55	0	0.0000	0.0000	0.0004	0.0000
-0.55	-0.50	1	0.0029	0.0001	0.0016	0.0001
-0.50	-0.45	1	0.0029	0.0003	0.0057	0.0002
-0.45	-0.40	5	0.0143	0.0010	0.0179	0.0007
-0.40	-0.35	29	0.0829	0.0051	0.0510	0.0024
-0.35	-0.30	44	0.1257	0.0114	0.1311	0.0066
-0.30	-0.25	88	0.2514	0.0240	0.3047	0.0170
-0.25	-0.20	238	0.6800	0.0580	0.6350	0.0397
-0.20	-0.15	432	1.2340	0.1197	1.1740	0.0840
-0.15	-0.10	678	1.9371	0.2166	1.8985	0.1603
-0.10	-0.05	918	2.6229	0.3477	2.6454	0.2743
-0.05	-0.00	1087	3.1057	0.5030	3.1347	0.4205
0.00	0.05	1123	3.2086	0.6634	3.1347	0.5795
0.05	0.10	934	2.6686	0.7969	2.6454	0.7257
0.10	0.15	677	1.9343	0.8936	1.8985	0.8397
0.15	0.20	400	1.1429	0.9507	1.1739	0.9160
0.20	0.25	199	0.5686	0.9791	0.6350	0.9603
0.25	0.30	90	0.2571	0.9920	0.3047	0.9830
0.30	0.35	35	0.1000	0.9970	0.1311	0.9934
0.35	0.40	16	0.0457	0.9993	0.0510	0.9976
0.40	0.45	3	0.0143	1.0000	0.0179	0.9992
0.45	0.50	0	0.0000	1.0000	0.0057	0.9997
0.50	0.55	0	0.0000	1.0000	0.0016	0.9999
0.55	0.60	0	0.0000	1.0000	0.0004	1.0000
0.60	0.65	0	0.0000	1.0000	0.0001	1.0000
0.65	0.70	0	0.0000	1.0000	0.0000	1.0000
0.70	0.75	0	0.0000	1.0000	0.0000	1.0000
0.75	0.80	0	0.0000	1.0000	0.0000	1.0000
0.80	0.85	0	0.0000	1.0000	0.0000	1.0000
0.85	0.90	0	0.0000	1.0000	0.0000	1.0000
0.90	0.95	0	0.0000	1.0000	0.0000	1.0000
0.95	1.00	0	0.0000	1.0000	0.0000	1.0000

Table 5.6

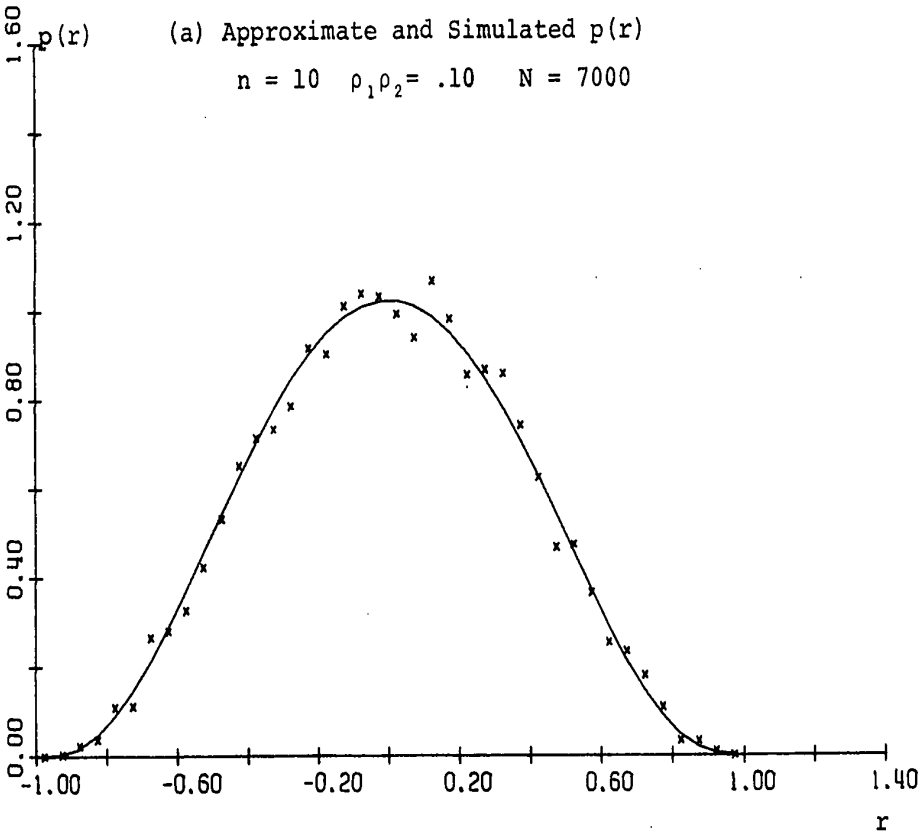
Simulated and Approximate Distributions of Cross-Correlation
N = 7000 n = 30 $\rho_1\rho_2$ = -.10

Range		Freq.	Gen.P(r)	Gen.CDF	Appr.P(r)	Appr.CDF
-1.00	-0.95	0	0.0000	0.0000	0.0000	0.0000
-0.95	-0.90	0	0.0000	0.0000	0.0000	0.0000
-0.90	-0.85	0	0.0000	0.0000	0.0000	0.0000
-0.85	-0.80	0	0.0000	0.0000	0.0000	0.0000
-0.80	-0.75	0	0.0000	0.0000	0.0000	0.0000
-0.75	-0.70	0	0.0000	0.0000	0.0000	0.0000
-0.70	-0.65	0	0.0000	0.0000	0.0000	0.0000
-0.65	-0.60	2	0.0057	0.0003	0.0014	0.0000
-0.60	-0.55	5	0.0143	0.0010	0.0053	0.0002
-0.55	-0.50	3	0.0086	0.0014	0.0167	0.0007
-0.50	-0.45	12	0.0343	0.0031	0.0450	0.0021
-0.45	-0.40	40	0.1143	0.0089	0.1052	0.0057
-0.40	-0.35	88	0.2514	0.0214	0.2169	0.0135
-0.35	-0.30	130	0.3714	0.0400	0.3999	0.0286
-0.30	-0.25	245	0.7000	0.0750	0.6661	0.0549
-0.25	-0.20	337	0.9629	0.1231	1.0093	0.0965
-0.20	-0.15	479	1.3686	0.1916	1.3992	0.1566
-0.15	-0.10	618	1.7657	0.2799	1.7818	0.2363
-0.10	-0.05	701	2.0029	0.3800	2.0901	0.3336
-0.05	-0.00	846	2.4171	0.5009	2.2628	0.4431
0.00	0.05	733	2.0943	0.6056	2.2628	0.5569
0.05	0.10	757	2.1629	0.7137	2.0901	0.6664
0.10	0.15	664	1.8543	0.8064	1.7817	0.7637
0.15	0.20	448	1.3943	0.8761	1.3992	0.8434
0.20	0.25	322	0.9200	0.9221	1.0093	0.9034
0.25	0.30	261	0.7457	0.9594	0.6661	0.9451
0.30	0.35	154	0.4400	0.9814	0.3999	0.9714
0.35	0.40	70	0.2000	0.9914	0.2169	0.9864
0.40	0.45	30	0.0857	0.9957	0.1052	0.9942
0.45	0.50	20	0.0571	0.9986	0.0450	0.9979
0.50	0.55	7	0.0200	0.9996	0.0167	0.9993
0.55	0.60	2	0.0057	0.9999	0.0053	0.9998
0.60	0.65	1	0.0028	1.0000	0.0013	1.0000
0.65	0.70	0	0.0000	1.0000	0.0003	1.0000
0.70	0.75	0	0.0000	1.0000	0.0000	1.0000
0.75	0.80	0	0.0000	1.0000	0.0000	1.0000
0.80	0.85	0	0.0000	1.0000	0.0000	1.0000
0.85	0.90	0	0.0000	1.0000	0.0000	1.0000
0.90	0.95	0	0.0000	1.0000	0.0000	1.0000
0.95	1.00	0	0.0000	1.0000	0.0000	1.0000

Fig.5.1

(a) Approximate and Simulated p(r)

n = 10 $\rho_1\rho_2$= .10 N = 7000



(b) Approximate and Simulated F(r)

n = 10 $\rho_1\rho_2$= .10 N = 7000

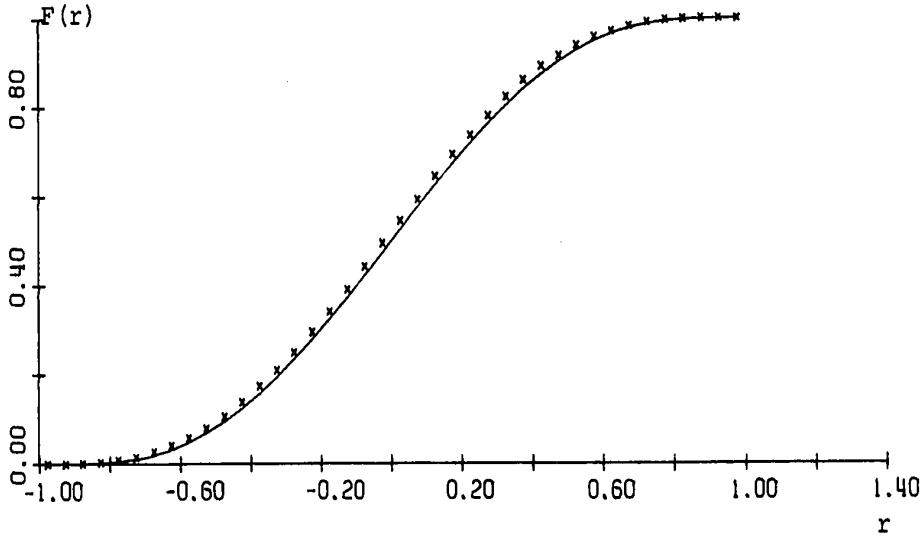
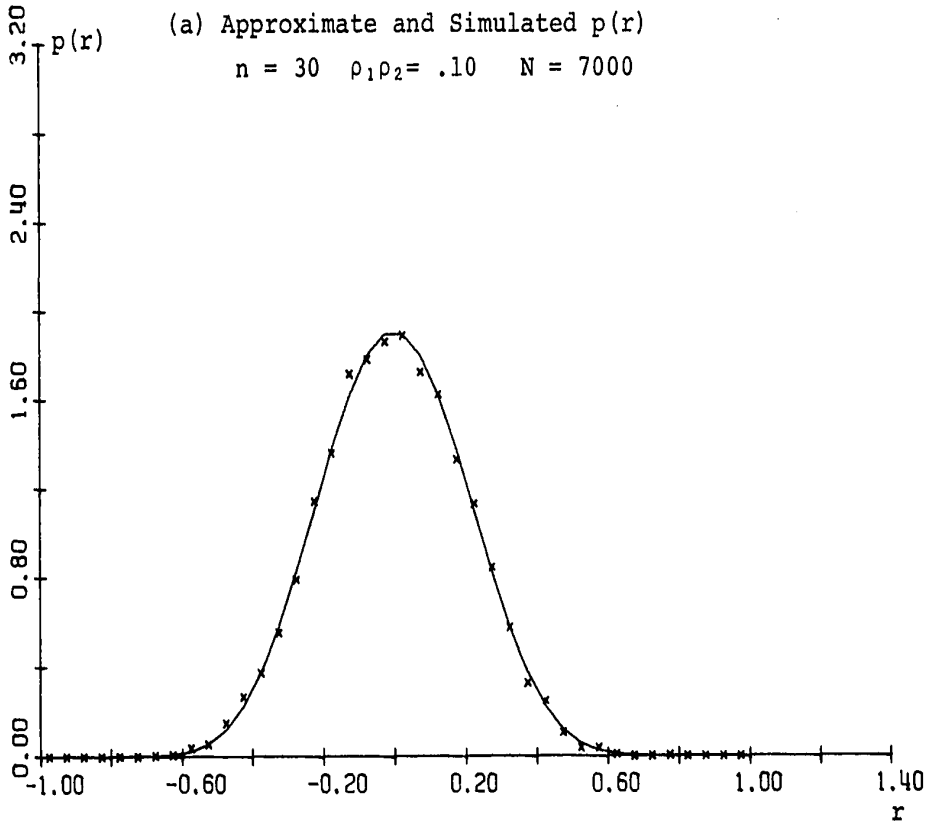


Fig.5.2

(a) Approximate and Simulated p(r)

n = 30 $\rho_1\rho_2$= .10 N = 7000



(b) Approximate and Simulated F(r)

n = 30 $\rho_1\rho_2$= .10 N = 7000

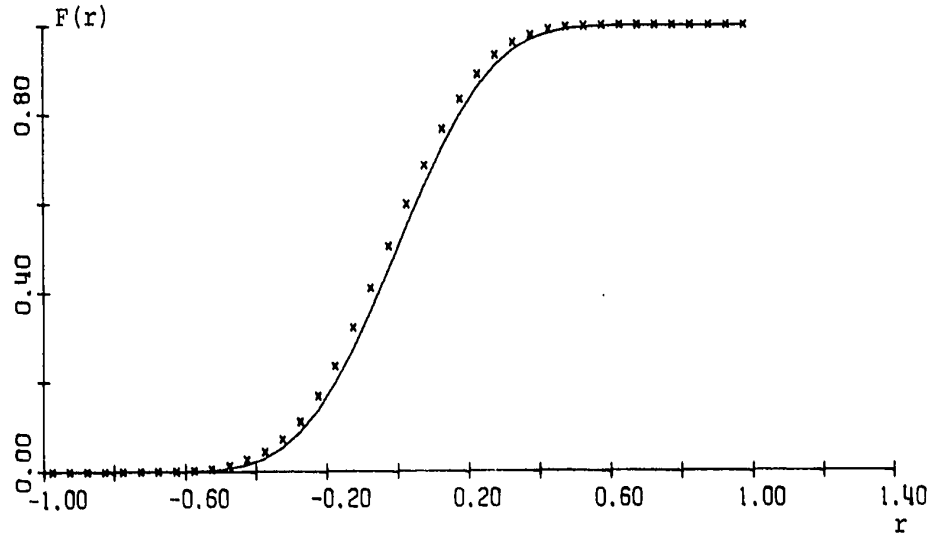
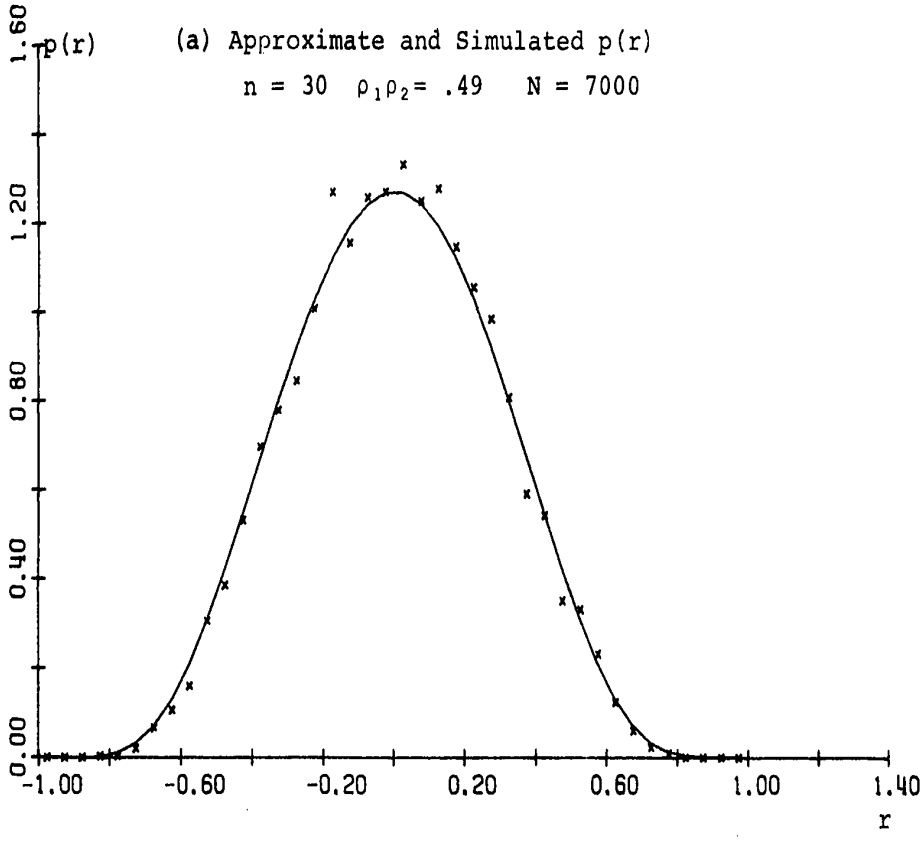


Fig.5.3

(a) Approximate and Simulated p(r)

n = 30 $\rho_1\rho_2$= .49 N = 7000



(b) Approximate and Simulated F(r)

n = 30 $\rho_1\rho_2$= .49 N = 7000

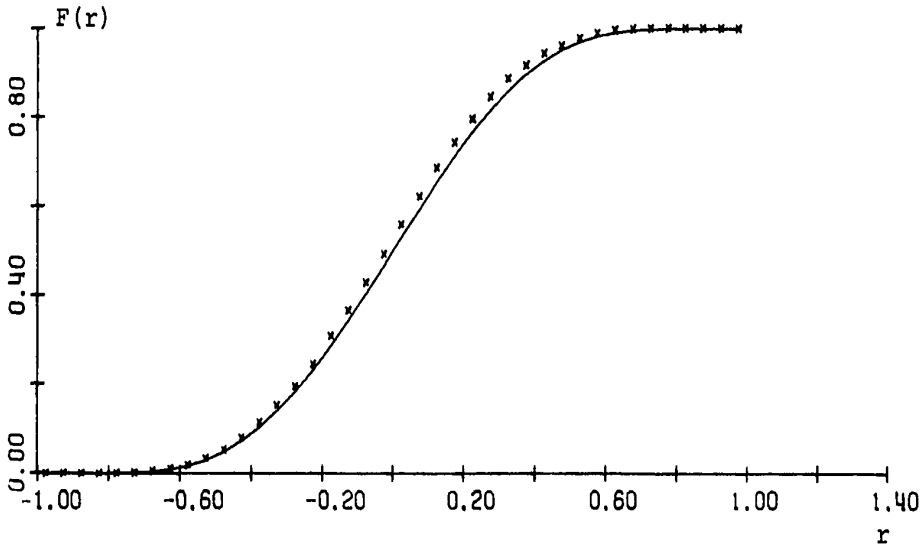
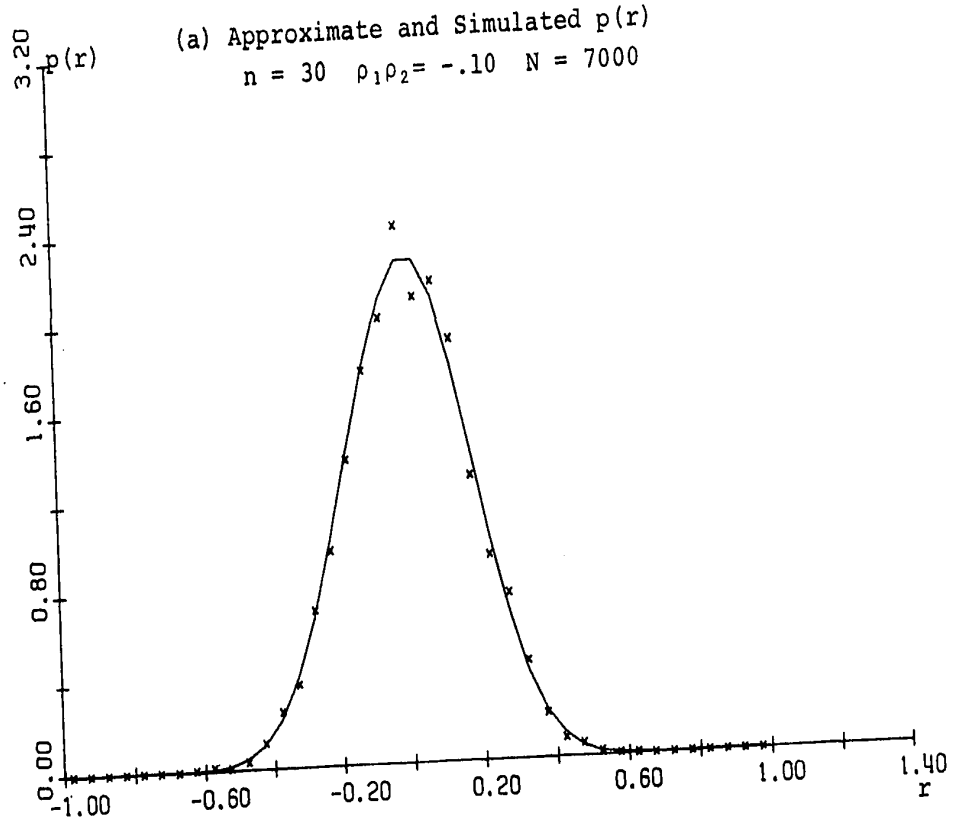


Fig.5.4

(a) Approximate and Simulated p(r)

n = 30 $\rho_1\rho_2$= -.10 N = 7000



(b) Approximate and Simulated F(r)

n = 30 $\rho_1\rho_2$= -.10 N = 7000

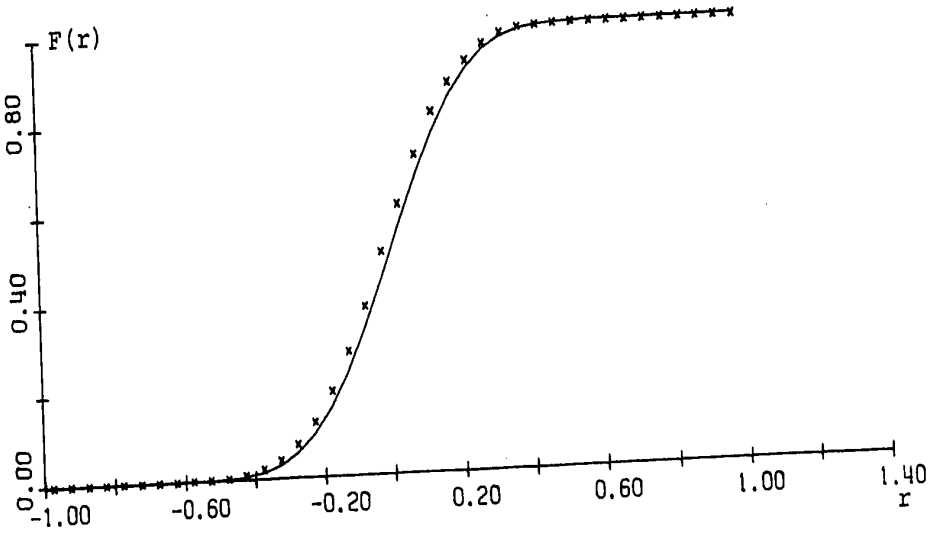
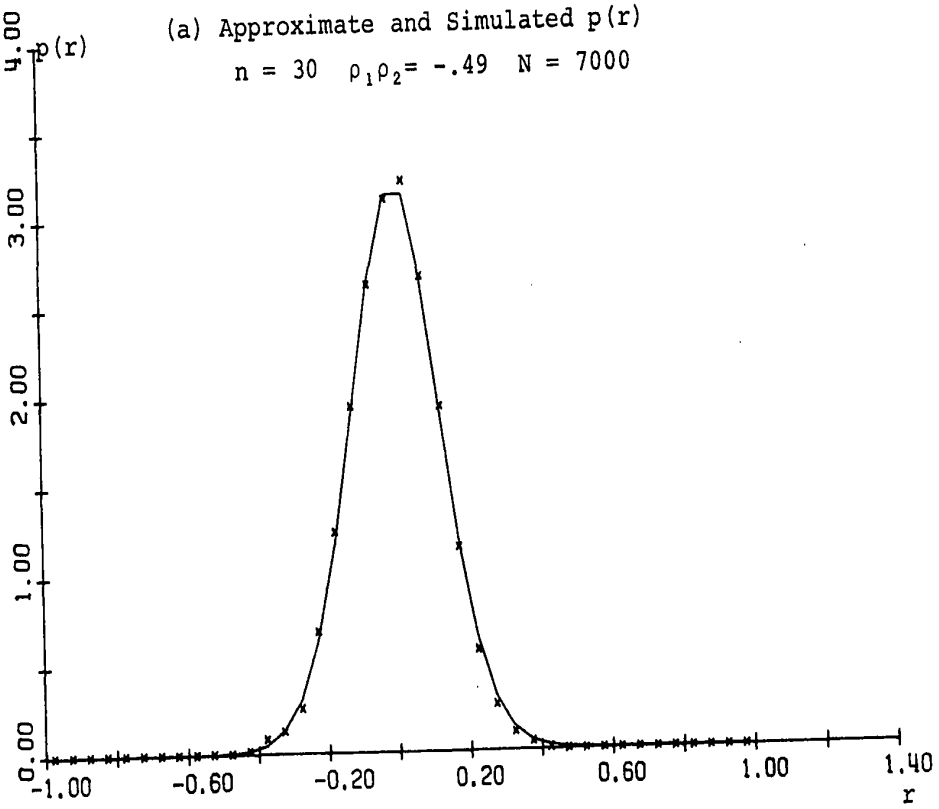


Fig.5.5

(a) Approximate and Simulated p(r)

n = 30 $\rho_1\rho_2$= -.49 N = 7000



(b) Approximate and Simulated F(r)

n = 30 $\rho_1\rho_2$= -.49 N = 7000

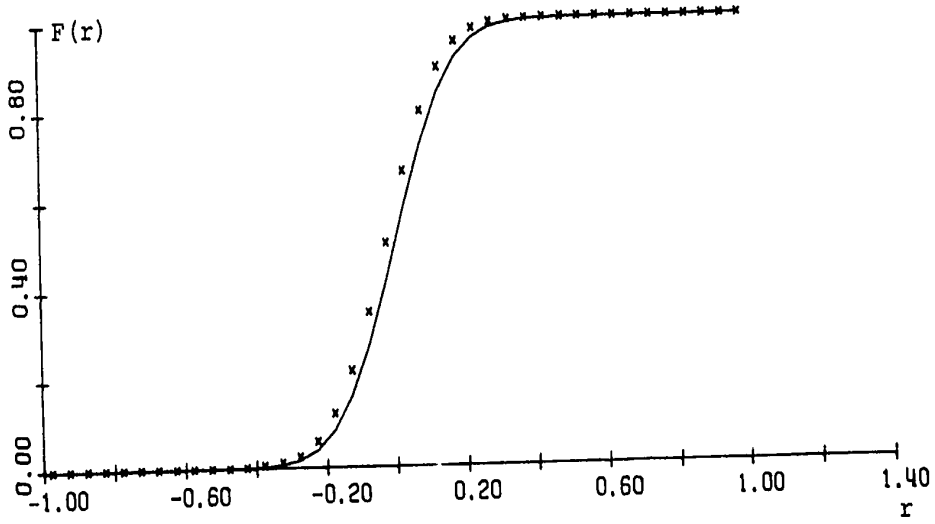
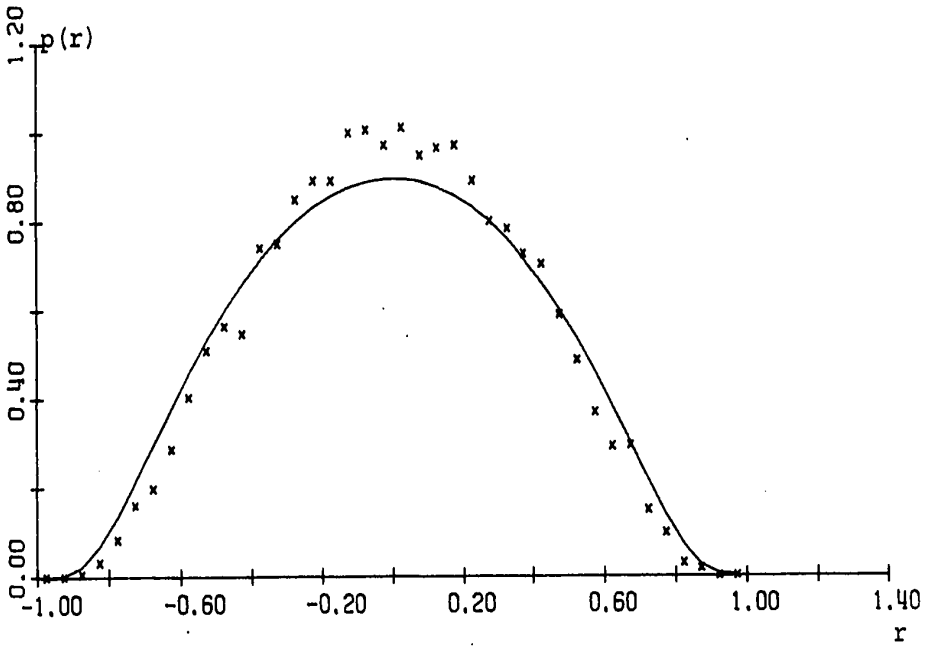


Fig.5.6

(a) Approximate and Simulated p(r)

n = 30 $\rho_1\rho_2$= .72 N = 7000



(b) Approximate and Simulated p(r)

n = 30 $\rho_1\rho_2$= .81 N = 7000

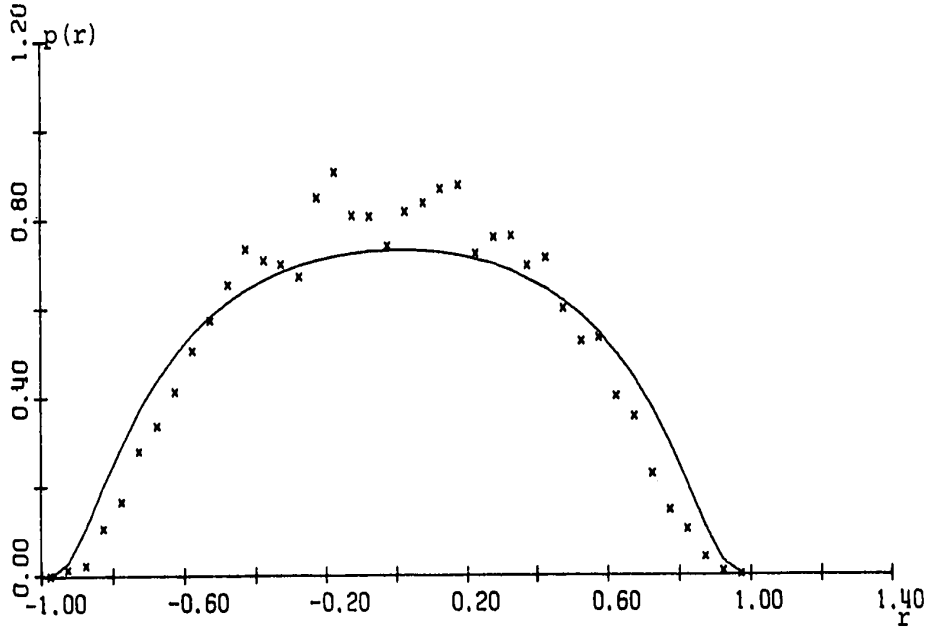
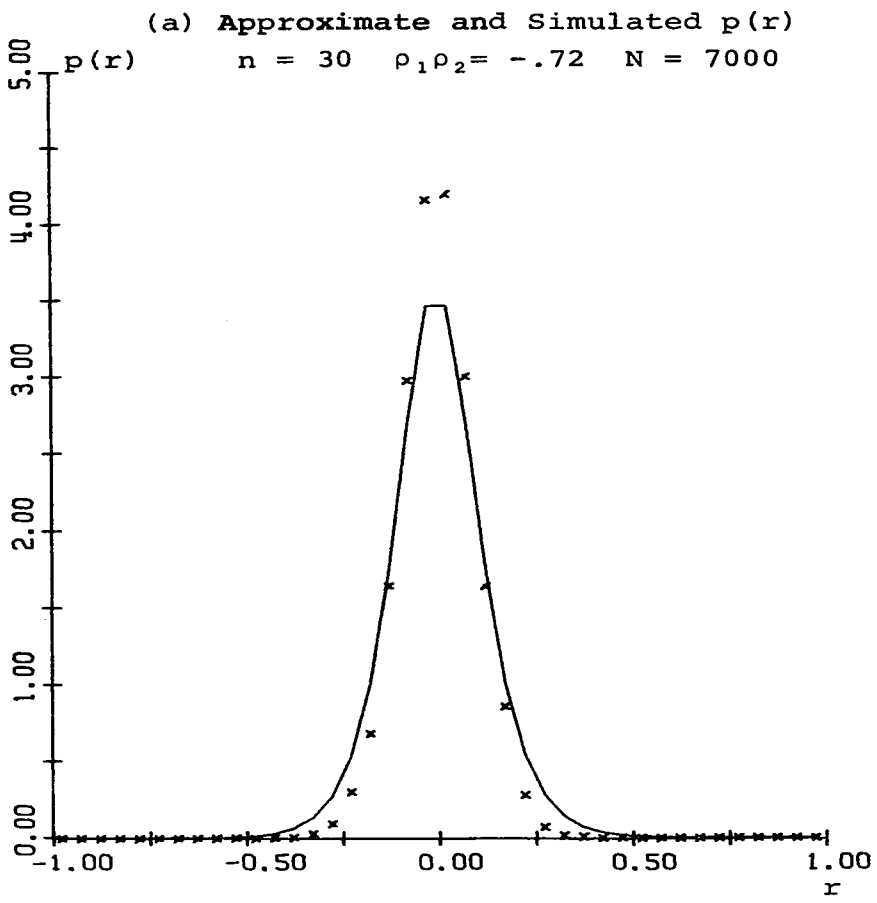


Fig.5.7


(a) Approximate and Simulated p(r)
p(r)
n = 30 $\rho_1\rho_2$= -.72 N = 7000
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-0.50
0.00
0.50
1.00
r

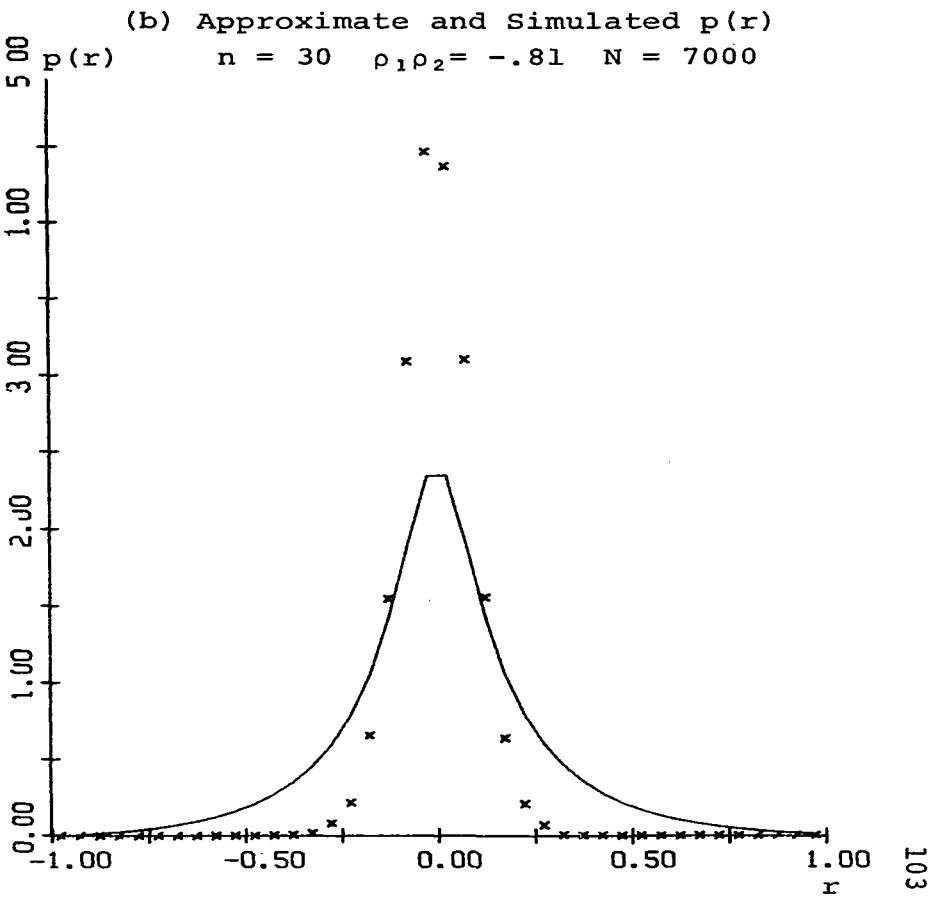

(b) Approximate and Simulated p(r)
p(r)
n = 30 $\rho_1\rho_2$= -.81 N = 7000
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-0.50
0.00
0.50
1.00
r

Table 5.7

Moments, Skewness, Kurtosis of Simulated and Approximate Distribution.

$\rho_1\rho_2 = .1$ N = 7000

	n = 10		n = 30	
	Appr.	Sim.	Appr.	Sim.
Mean	0.0000	0.0052	0.0000	-.0027
Variance	0.1213	0.1217	0.0405	0.0408
Third Moment	0.0000	-.0003	0.0000	-.0001
Fourth Moment	0.0349	0.0356	0.0045	0.0046
Skewness	0.0000	-.0078	0.0000	0.0012
Kurtosis	2.3800	2.4038	2.7480	2.7761

Table 5.8

Moments, Skewness, Kurtosis of Simulated and Approximate Distribution.

n = 30 N = 7000

	$\rho_1\rho_2$ =-.10		$\rho_1\rho_2$ =-.49		$\rho_1\rho_2$ =-.72		$\rho_1\rho_2$ =-.81	
	Appr.	Sim.	Appr.	Sim.	Appr.	Sim.	Appr.	Sim.
Mean	0.000	0.001	0.000	-.009	0.000	0.001	0.000	0.000
Var.	0.029	0.029	0.016	0.015	0.019	0.009	0.059	0.008
3rd Mom.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4th Mom.	0.002	0.002	0.001	0.001	0.001	0.000	0.015	0.000
Skew	0.000	-.025	0.000	-.059	0.000	-.053	0.000	-.048
Kurt.	2.865	2.899	3.186	3.122	3.908	3.457	4.306	3.450

Table 5.9

Moments, Skewness, Kurtosis of Simulated and Approximate Distribution.

n = 30 N = 7000

	$\rho_1\rho_2$ = .49		$\rho_1\rho_2$ = .72		$\rho_1\rho_2$ = .81	
	Appr.	Sim.	Appr.	Sim.	Appr.	Sim.
Mean	0.000	0.003	0.000	0.006	0.000	-.003
Var.	0.082	0.078	0.143	0.123	0.190	0.156
3rd Mom.	0.000	0.000	0.000	0.000	0.000	0.000
4th Mom.	0.017	0.015	0.045	0.035	0.074	0.052
Skew.	0.000	-.002	0.000	0.001	0.000	0.003
Kurt.	2.470	2.492	2.200	2.301	2.040	2.133

The values of $D_{SA}(N)$ obtained for $\rho_1\rho_2$ = .1, .49 and n=10, 30 are well below the .01 significant point, $D_{.01}(7000)$ = .0172 (and therefore less than .02). Hence, we can say with 99% confidence that the approximate distribution is accurate to within ± .04 of the true distribution for these values of the parameters. However, for higher values of $\rho_1\rho_2$ the results are not as satisfactory. For $\rho_1\rho_2$ = .72, .81, n = 30, the values of $D_{SA}(N)$ exceed .02 indicating a possible error in the approximate distribution greater than .04. A comparison of the simulated and approximate distributions in Fig. 5.7 and the values of the moments in Table 5.9 indicates the variance of the approximate distribution for high values of $\rho_1\rho_2$ and small n

(n = 30) is too large. Fig. 5.7 shows the simulated distribution has a narrower spread and higher peak than the approximate distribution. The difference between the two distributions is less severe for the lower value of $\rho_1\rho_2$ ($\rho_1\rho_2$= .72). From these observations it would appear that for high values of the autocorrelations, a more accurate distribution may be obtained only with larger values of n, since the variance of the approximate distribution decreases with increasing n. This can be observed from the graphs in Fig. 1.1. The approximate distribution for $\rho_1\rho_2$ = .72 has a variance of .0966 for n = 50 as compared to a variance of .1425 for n = 30.

Comparing the simulated and approximate distribution for negative products of the serial correlations (that is, $\rho_1\rho_2 < 0$), it can be seen from Table 5.1 and Figs. 5.4, 5.5 that the approximate distribution closely resembles the simulated one for $-.5 < \rho_1\rho_2 < 0$. However, for $\rho_1\rho_2$ = -.72, -.81 in Figs. 5.8, 5.9 the approximate distribution again seems to have a larger variance than the simulated distribution, where the simulated distribution (that is, the empirical frequency distribution) represents the true (unknown) distribution of the sample crosscorrelation r_{XY}. Thus the approximate distribution not only has larger variance for large positive $\rho_1\rho_2$ values but also indicates an increase in variance for decreasing values of $\rho_1\rho_2 < -.50$, which does not agree with the behaviour of the simulated distribution.

In order to illustrate this discrepancy in the two distributions the variances for both the approximate and the simulated distribution were plotted for $n = 30$ in Fig. 5.8 . The corresponding values of the variances are listed in Table 5.10. For this case, $n = 30$, the variance, $\sigma^2_{r_{XY}}$, for the approximate distribution appears to have a minimum close to $\rho_1\rho_2 = -.6$ whereas the simulated variance tends to zero as $\rho_1\rho_2$ tends to -1.

Table 5.10

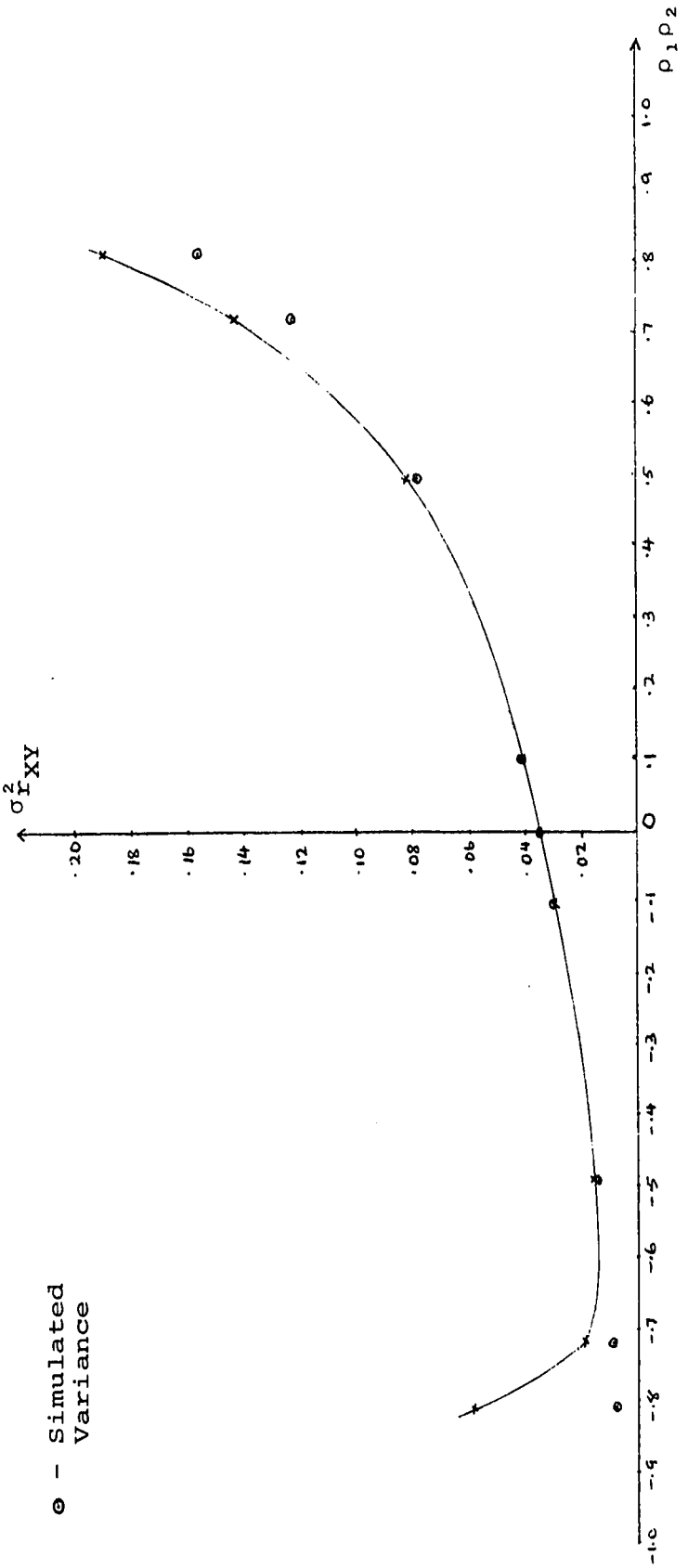
Variances of Approximate and Simulated Distribution of r_{XY} .

$n = 30$, $N = 7000$

$\rho_1\rho_2$	Variance	
	Approx.	Sim.
-.81	.059	.008
-.72	.019	.009
-.49	.016	.015
-.10	.029	.030
.00	.035	.035
.10	.041	.041
.49	.082	.078
.72	.143	.123
.81	.190	.156

Fig.5.8

Variance of Approximate and Simulated Distribution against $\rho_1\rho_2$



The primary implication of the results given in this section appears to be the following. The approximate distribution for values of $-.5 \leq \rho_1\rho_2 \leq .5$ and $n \leq 30$ is accurate to within $\pm .04$ of the true distribution of cross-correlation with 99% probability. For high values of $|\rho_1\rho_2|$ and the same values of n, the approximate distribution appears to have too large a variance. The error in the distribution in this case could be greater than .04. For larger values of n, however, the approximate distribution for high values of $|\rho_1\rho_2|$ could be a more accurate representation of the true distribution, since the variance of the distribution decreases with increasing n.

To obtain a more definite pattern of behaviour of the approximate distribution with respect to the parameters, n, $\rho_1\rho_2$, more thorough testing would have to be performed on the distribution for various combinations of the parameters. Due to the high computer cost involved in simulating and testing each distribution (see Table C in Appendix C) we have limited the testing of the approximate distribution to the few values of the parameters considered above, and observations on the behaviour of this distribution have been made based on the results of these tests.

Since the degree of accuracy in the approximate distribution for a particular value of $\rho_1\rho_2$ varies with n, it is possible to test for a value of n which will give a desired

precision in the approximate distribution for that value of $\rho_1\rho_2$. For practical purposes, however, it would be useful to establish for each value of $\rho_1\rho_2$, a minimum value of n with which the approximate distribution may be applied with reasonable accuracy (since in practice, series of data encountered are usually small). In the next section we shall apply the confidence band technique to estimate a minimum value of n for a particular value of $\rho_1\rho_2$.

5.3 Estimation of a Minimum Value for Sample Size.

Since the approximate density function as given by Eqn. (1.5) is dependent only on the sample size n and the product of the autocorrelations $\rho_1\rho_2$, it is useful to be able to establish a minimum value of n, say n_m , for which the approximate distribution for a particular value of $\rho_1\rho_2$ can be used with reasonable accuracy.

Let ε be the desired accuracy we wish to have in the approximate distribution. Let N_ε be the sample size which will give a simulated distribution within $\pm\varepsilon/2$ of the true distribution with 99% probability. The algorithm for estimating n_m , given a value of $\rho_1\rho_2$, may be summarized in the following steps :

1) For an initial value of n, evaluate $F_A(r;n,k)$ using Eqn. (1.5).

2) Simulate the distribution $F_S(r;n,k)$ using a sample of size N_ε.

3) Compute $D_{SA}(N_\varepsilon;n,k)$ by Eqn. (5.4).

4) If $D_{SA}(N_\varepsilon;n,k) < \varepsilon$, put $n = n - 1$ and go to (1).
If $D_{SA}(N_\varepsilon;n,k) \geq \varepsilon$, put $n_m = n - 1$ and exit.

Taking $\varepsilon = .04$, $N_\varepsilon = 7000$, the value of n_m estimated by the above algorithm was 6 for $\rho_1\rho_2 = .1$ and 10 for $\rho_1\rho_2 = .49$. The above algorithm can be used to determine the smallest sample size that can be used with each value of $\rho_1\rho_2$ in $|\rho_1\rho_2| < 1$. However, from the discussion in Section 5.2 the range of $\rho_1\rho_2$- values for which these calculations should be applied is (-.50,.50).

5.4 Critical Values of the Approximate Distribution.

Approximate critical values of r_{XY} for $\rho_1\rho_2 \neq 0$ can be determined using the approximate density function of Eqn. (1.5). It is, however, necessary to estimate an error bound for these approximate critical values. The procedure used is based on the theory on error bounds described in Section 4.6.

We first simulate the distribution of r for $\rho_1\rho_2 = k \neq 0$ using a sample size of 7000 for the given value of n, and evaluate the q-critical value, $r_{S,q}(n,k)$ (refer Section 4.9). Then we can say with $(1-\alpha)100\%$ certainty that

$$|r_{T,q}(n,k) - r_{S,q}(n,k)| < u_{N,q,\alpha} \qquad (5.5)$$

where $r_{T,q}(n,k)$ is unknown and $u_{N,q,\alpha}$ is the critical value of $u_{N,q}$ given by Eqn. (4.23).

Note from Eqn. (4.23) that, to compute $u_{N,q,\alpha}$, we require the value of the true density function at the critical point. However, since both these values are unknown and since we have shown that the simulated distribution has a good fit at the tail regions for $\rho_1\rho_2$ in (-.5,.5), we shall replace $p_T(r_{T,q};n,k)$ by the simulated value, that is, we compute

$$u_{N,q,\alpha} = z_\alpha(pq)^{1/2} / [N^{1/2}p_S(r_{S,q};n,\rho_1\rho_2)] \quad , \qquad (5.6)$$

Next, we evaluate the approximate critical value, $r_{A,q}(n,k)$, using the approximate density function of Eqn. (1.5), and then compute the absolute difference between the approximate and simulated critical values,

$$|r_{A,q}(n,k) - r_{S,q}(n,k)| \quad . \tag{5.7}$$

If $|r_{A,q}(n,k) - r_{S,q}(n,k)| < u_{N,q,\alpha}$, then we know that

$$\begin{aligned} |r_{A,q}(n,k) - r_{T,q}(n,k)| &= |r_{A,q}(n,k) - r_{S,q}(n,k) - r_{T,q}(n,k) + r_{S,q}(n,k)| \\ &< |r_{A,q}(n,k) - r_{S,q}(n,k)| \\ &\quad + |r_{T,q}(n,k) - r_{S,q}(n,k)| \\ &< 2u_{N,q,\alpha} \end{aligned} \tag{5.8}$$

Hence, $2u_{N,q,\alpha}$ can be used as an approximate error bound for the critical value computed from the approximate distribution. That is, we can say with $(1-\alpha)100\%$ certainty that the true critical value is given by

$$r_{T,q}(n,k) = r_{A,q}(n,k) \pm 2u_{N,q,\alpha} \tag{5.9}$$

To illustrate the above method, the critical values of the simulated and approximate distributions of r for $\rho_1\rho_2 = .1$,

$n = 10, 30$ and $N = 7000$, and values of $u_{N,q,\alpha}$ for $\alpha = .01$ are computed and shown in Table 5.11. In all cases, the difference (5.7) is less than the corresponding $u_{N,q,.01}$. Hence, we may conclude with 99% certainty that the true critical value for, say, $q = .01$ and $n = 10$ is given by

$$r_{T,.01}(10,.1) = r_{A,.01}(10,.1) \pm .0332$$
$$= .7340 \pm .0332 .$$

Critical values of the simulated and approximate distribution of r_{XY} for $\rho_1\rho_2 = -.49, -.1, .49$, $n = 30$ and $N = 7000$ are shown in Table 5.12.

Having devised methods of estimating the degree of accuracy obtainable with the approximate distribution for various combinations of the parameters $(n, \rho_1\rho_2)$ and approximating the error in the critical points computed from this distribution, it is now possible to apply the approximate distribution in tests of significance for correlation. Since statistical tests have indicated good accuracy in the approximate distribution, for the small n values and $\rho_1\rho_2$ in $(-.5,.5)$, it may be reasonable to use critical values, for these values of the parameters, evaluated from this distribution as the exact values, assuming the errors in them to be neglible. A table of critical values of r computed from the approximate distribution for various values of n and $\rho_1\rho_2$ is given in Appendix Table D2.

Table 5.11

Approximate and Simulated Critical Values of r_{XY} and the Corresponding $u_{N,q,\alpha}$ Values.

$N = 7000$, $\alpha = .01$

$\rho_1\rho_2$	n	q	$r_{A,q}$	$r_{S,q}$	$\lvert r_{A,q} - r_{S,q} \rvert$	$u_{N,q,.01}$
.1	10	.01	.7340	.7372	.0032	.0166
		.02	.6756	.6780	.0024	.0166
		.05	.5724	.5737	.0013	.0163
	30	.01	.4545	.4527	.0018	.0169
		.02	.4069	.4089	.0019	.0140
		.05	.3319	.3301	.0019	.0112

Table 5.12

Approximate and Simulated Critical Values of r_{XY} .

$N = 7000$, $\alpha = .01$

$\rho_1\rho_2$	n	q	$r_{A,q}$	$r_{S,q}$	$\lvert r_{A,q} - r_{S,q} \rvert$
-.10	30	.010	.3930	.3952	.0022
		.025	.3340	.3361	.0021
		.050	.2830	.2880	.0050
-.49	30	.010	.3040	.3128	.0088
		.025	.2530	.2537	.0007
		.050	.2100	.2129	.0029
.49	30	.010	.6163	.6083	.0080
		.025	.5433	.5428	.0002
		.050	.4711	.4685	.0026

CHAPTER VI

APPLICATION, CONCLUSION AND DISCUSSIONS.

6.1 Application of the Approximate Distribution of Cross-Correlation.

We shall present in this section an example to illustrate a practical application of the approximate distribution of the cross-correlation in economic time series analysis. The example shows that using the approximate distribution it is possible to test for correlation between two processes (that is, two series of data) without having to assume that their serial correlations are zero.

The time series dealt with here is the forecast errors in the interest rates on the U.S. Federal Fund given in a paper by Janssen [16]. Forecast errors for Friday, Monday, Tuesday and Wednesday were observed over a period of 33 weeks. The following notation was used :

t : time index, Thursday (t=1), Friday (t=2), Monday (t=3), Tuesday (t=4), Wednesday (t=5)

v_t : forecast error on day t

The forecast errors appear in Table 6.1.

Table 6.1

Forecast Errors for the Period 4.1.68 - 30.10.68
(33 Weeks)

Fri	Mon	Tue	Wed
2	3	4	5
-0.007	0.000	-09500	-1.427
-0.199	0.111	0.259	-1.096
0.081	-0.117	-0.383	0.469
-0.477	0.215	0.267	0.286
-0.132	-0.258	-0.019	0.133
0.011	0.003	0.001	-0.109
0.010	0.128	-0.364	-0.392
0.306	0.025	0.128	0.288
-0.142	-0.008	-0.124	-0.362
0.047	0.131	0.137	0.170
-0.018	0.001	-0.498	-0.767
-0.042	-0.126	-0.258	0.122
0.367	0.030	-0.371	-0.136
-0.077	0.122	0.136	-0.208
-0.116	-0.131	0.117	0.534
-0.016	0.127	-0.739	-0.413
-0.915	0.558	-0.582	0.092
0.492	0.039	0.129	0.035
-0.009	0.252	0.146	0.043
-0.018	0.001	-0.123	0.264
-0.007	-0.123	-0.133	0.260
0.118	0.011	0.127	-0.087
0.000	0.002	0.127	-0.087
0.000	-0.123	-0.008	0.022
-0.124	-0.007	0.126	0.157
-0.143	0.009	0.000	0.391
-0.007	0.002	-0.123	-0.740
0.045	0.005	-0.124	-0.111
0.125	0.136	0.136	0.154
0.002	-0.498	0.463	0.181
-0.133	0.117	-0.115	0.261
0.118	0.011	-0.123	-0.611
-0.089	-0.005	0.126	0.404

In his analysis, Janssen used the t-distribution to test the hypothesis that no correlation exists between forecast errors of two different days, that is,

$$\rho(v_t , v_s) = 0 \qquad (6.1)$$

for $t = 2, 3, 4$ and $s = 3, 4, 5$, $t \neq s$. The t-test for correlation requires that at least one of the variables be normally distributed. This condition is satisfied in this case since the author has established, in an earlier part of his analysis, the distribution of the forecast errors to be approximately normal by hypothesis testing. The coefficients of correlation between each of the four series, and the associated t-values, as obtained by Janssen, are shown in Table 6.2. Critical values of the t-distribution and the associated critical correlation coefficients are :

$$\left.\begin{array}{lll} t_{90} = 1.31 & \text{or} & r_{v_t v_s, .10} = .229 \\ t_{95} = 1.70 & \text{or} & r_{v_t v_s, .05} = .292 \\ t_{99.5} = 2.75 & \text{or} & r_{v_t v_s, .005} = .443 \end{array}\right\} \text{for all } s, t$$

where $r_{v_t v_s, \alpha}$ is the α-critical point of the sample cross-correlation $r_{v_t v_s}$ (that is, $P[r_{v_t v_s} \leq r_{v_t v_s, \alpha}] = 1-\alpha$), and a value of $r_{v_t v_s}$ is denoted by $r(v_t, v_s)$.

Janssen observed that the correlations between successive

days are large enough to cause rejection of the hypothesis (6.1), that is, rejection of the assumption of independence between these series of forecast errors. However, for longer lags (that is, not adjacent days) the results do not give reasons to reject the hypothesis.

Table 6.2

Janssen's Observed Coefficients of Correlation, $r(v_t, v_s)$ and Associated t-test Values between Series of Forecast Errors (Days).

Time	t \ s	Mon. 3	Tue. 4	Wed. 5
Fri.	2	-.425 (2.61)	.131 (.736)	-.052 (.290)
Mon.	3		-.280 (1.63)	-.131 (.735)
Tue.	4			.345 (2.05)

We shall perform here the test for correlation between each of the four error series using the approximate distribution (Eqn. (1.5)). To apply this distribution the autocorrelation in each series must be known. The autocorrelation of lag 1 for each of the error series can be estimated from the series of

observations in Table 6.1 using the formula (see Jenkins and Watts [18]),

$$r(v_t;1) = \frac{\sum_{i=1}^{n-1}(v_{ti} - \bar{v}_t)(v_{ti(i+1)} - \bar{v}_t)}{\sum_{i=1}^{n-1}(v_{ti} - \bar{v}_t)^2} , \qquad (6.2)$$

where $\bar{v}_t = \frac{1}{n}\sum_{i=1}^{n} v_{ti}$.

Estimates of the autocorrelation of lag 1 for each of the error series are shown in Table 6.3. In order to be able to apply the approximate distribution we shall assume that the true autocorrelations (serial correlations) are given to be

$\rho_2 = -.3133$, $\rho_3 = -.1752$, $\rho_4 = -.0532$, $\rho_5 = -.0192$,

where ρ_j is a given serial correlation for the j^{th} series of forecast errors.

Table 6.3

Estimates of Autocorrelation of Lag 1

	t	$r(v_t;1)$
Fri.	2	-.3133
Mon.	3	-.1752
Tue.	4	-.0532
Wed.	5	-.0192

The sample cross-correlations between each of the four error series are estimated by the following formula, using a sample size of 33,

$$r(v_t, v_s) = \frac{\sum_{i=1}^{33} (v_{ti} - \bar{v}_t)(v_{si} - \bar{v}_s)}{[\sum_{i=1}^{33} (v_{ti} - \bar{v}_t)^2 \sum_{i=1}^{33} (v_{si} - \bar{v}_s)^2]^{1/2}} \qquad (6.3)$$

for $t = 2, 3, 4$, $s = 3, 4, 5$ and $t \neq s$. These estimates which are the same as those obtained by Janssen in Table 6.2 are shown in Table 6.4.

Table 6.4

Sample Cross-Correlation between Error Series

Time	t \ s	Mon. 3	Tue. 4	Wed. 5
Fri.	2	-.425	.131	-.052
Mon.	3		-.280	-.131
Tue.	4			.345

To test the null hypothesis that

$$\rho\ (v_t\ ,\ v_s) \ = \ 0$$

for $t = 2, 3, 4$, $s = 3, 4, 5$ and $t \neq s$, critical values of $r_{v_t v_s}$ must be available. The relevant values of $\rho_t \rho_s$ are shown in Table 6.5, where ρ_t and ρ_s are the specified values given earlier. Using the approximate distribution in Eqn.(1.5) the α-critical values of $r_{v_t v_s}$ are computed for $\alpha = .005, .01, .025, .05, .10$, $n = 33$ and the relevant $\rho_t \rho_s$ values, and are tabulated in Table 6.6.

Table 6.5

Product of Autocorrelations of Lag 1

$\rho_t \rho_s$

t	s / ρ_s / ρ_t	3	4	5
		-.1752	-.0532	-.0192
2	-.3133	.0549		
3	-.1752	.0167	.0093	
4	-.0532	.0060	.0034	.0010

Table 6.6

One-Tail Critical Values of $r_{v_t v_s}$

n = 33

$\rho_t\rho_s$ \ α	Critical Values $r_{v_t v_s,\alpha}$				
	.005	.010	.025	.050	.100
.0549	.460	.420	.359	.305	.240
.0167	.447	.408	.349	.295	.232
.0093	.445	.406	.348	.293	.231
.0060	.444	.405	.346	.293	.230
.0034	.443	.404	.345	.292	.229
.0010	.442	.403	.344	.292	.229

For $\rho_2\rho_3 = .0549$ the observed cross-correlation $r(v_2, v_3) = -.425$ is less than the .025-critical point $-r_{v_2 v_3, .025} = -.359$ for a two-tail test; and, hence, the null hypothesis of $\rho(v_2, v_3) = 0$ must be rejected at the .05 level. This implies that the cross-correlation between the Friday and Monday forecast errors has a negative value and is not zero; hence, these two series of error forecasts are not independent. Similarly, for $\rho_4\rho_5 = .0010$, $r(v_4, v_5) = .345$ which exceeds $r_{v_4 v_5, .025} = .344$. This implies that the series of Wednesday forecast errors depends on the Tuesday forecast errors. The

other two adjacent days which indicate dependence in their forecast error series are Monday and Tuesday. Comparing the Friday forecast error series with those of Tuesday, $r(v_2, v_4) = .131 < r_{v_2 v_4, \alpha/2}$, and those of Wednesday $r(v_2, v_5) = -.052 > -r_{v_2 v_5, \alpha/2}$, their cross-correlations are not significantly different from zero, implying that these series are linearly independent. Similarly, the cross-correlation between the Monday and Wednesday forecast errors, $r(v_3, v_5) = -.131$, is not significantly different from zero.

These results are similar to those obtained by Janssen. In using the t-test Janssen made the assumption that the product of the serial correlations is zero (that is, $\rho_t \rho_s = 0$); or in other words that at least one of the series comes from a normal distribution and hence, $r_{v_t v_s}$ had the Pearson correlation coefficient distribution. In this particular example of forecast errors the serial correlations were relatively small, yielding even smaller products, see Table 6.5. Since, these products are all fairly close to zero, Janssen's assumption did not result in erroneous decision. Perhaps the Friday - Monday relationship should be considered in more detail. For testing (two-tail test) Janssen used the critical value

$$r_{v_2 v_3, .025} = r_{.025}(33;\ \rho_2 \rho_3 = 0) = .344 ,$$

whereas the critical value from the approximate distribution is

$$r_{v_2 v_3, .025} = r_{.025}(33;\ \rho_2 \rho_3 = .0549) = .359 .$$

Thus even for a small serial correlation product, such as $\rho_2 \rho_3 = .055$, there exists a noticiable difference in the

critical values. Suppose the observed cross-correlation $r(v_2, v_3)$ had turned out to be -.350, then Janssen would have concluded that the series are dependent, whereas the critical points in Table 6.6 would have lead to the conclusion of independence.

The development in Chapter 5 showed that the critical points of the approximate distribution closely represent the true critical points of the true cross-product correlation distribution as long as the two series have small serial correlations with the same or opposite sign. Thus for series with such serial correlations the critical, sample-crosscorrelation values should be determined as shown in Appendix A instead of just using the critical points of the Pearson correlation coefficient distribution.

Emphasizing this point further, consider two series of size n = 30 with serial correlation product of $\rho_1\rho_2 = -.49$, see Table 5.1. The $\alpha = .025$ - critical value, as found in Table 5.12 is

$$r_{.025}(n=30;\ \rho_1\rho_2 = -.49) = .253$$

whereas, under the assumption of $\rho_1\rho_2 = 0$ the critical point is

$$r_{.025}(n=30;\ \rho_1\rho_2 = 0) = .361 \ .$$

Hence, using .361 the null hypothesis, $\rho_{XY} = 0$, will not be rejected (that is, series are not independent) when it actually should be. This could of course result in serious errors in practical situations.

6.2 Conclusion and Discussions.

To test whether or not the correlation between two series is significantly different from zero, the distribution of the sample cross-correlation, r, and its critical values must be available. McGregor and Bielenstein [30] have derived an approximate null distribution (Eqn. (1.5)) of the cross-correlation for two series of the linear, stationary Markov type with known (serial correlations) autocorrelations of lag 1, ρ_1 and ρ_2. This approximate distribution depends only on the sample size, n, of the series and the product of the autocorrelations, $\rho_1\rho_2$. For the case where $\rho_1\rho_2 = 0$, this distribution reduces to the null distribution of the Pearson correlation coefficient (Eqn. (1.4)).

To compute critical values of r from the approximate distribution an algorithm has to be designed to evaluate the approximate density function, $p^*(r;n, \rho_1\rho_2)$ (Eqn. (1.5)). Several terms in the expression for this density function tend to present overflow and underflow problems, which must be dealt with in the algorithm, see Section A1 in Appendix A. The Guass-Legendre quadrature formula is used for the numerical integration of the density function.

In order to apply the approximate distribution of cross-correlation in any valid test for correlation it is necessary to

determine the accuracy of this distribution and its critical values. The best approach to this problem is simulation since the Markov series required for investigation into the problem are not readily available in a suitable form in practice.

A distribution of the cross-correlation is simulated by taking N observations of the sample cross-correlation between samples of size n of two series generated by two linear, stationary Markov processes of the form,

$$X_t = \rho_1 X_{t-1} + Z_t$$

$$Y_t = \rho_2 Y_{t-1} + Z'_t$$

where ρ_1, ρ_2 are known autocorrelations of lag one and Z_t, Z'_t are independent $N(0,1)$ random variables. To simulate realizations of the Markov processes the main requirements are a criterion to determine stationarity of the processes and a 'good' random number generator which will generate independent normal random numbers satisfying various statistical criteria of randomness and normality.

The process of searching for a random number generator for the simulation is focused on its properties satisfying normality and randomness, especially with respect to serial correlation since the simulated Markov series should be unrelated. It was decided after careful consideration, see Section 3.6 , that the generator proposed by Chen [7] would be most suitable for this simulation.

Before applying the simulated distribution, it is necessary

to determine how closely it represents the true distribution of cross-correlations. The Kolmogorov-Smirnov and Anderson-Darling tests are used to determine the goodness-of-fit of the simulated distribution for the case $\rho_1\rho_2 = 0$, using the null distribution of the Pearson correlation coefficient (Eqn. (1.4)) as the hypothesized true distribution. Using the Kolmogorov-Smirnov criterion, the error in simulation is estimated to be at most $\pm .02$ over the entire cumulative distribution of cross-correlation, with 99% probability for a sample of size N = 7000. Results of the Anderson-Darling test show that there is a good fit of the simulated to the theoretical distribution at the tail regions. A comparison between the simulated and theoretical critical values of r using Bahadur's theory on sample quantiles, see Section 4.8 , also indicates reasonably good accuracy in the simulation.

To determine the accuracy of the approximate distribution (Eqn. (1.5)) for $\rho_1\rho_2 \neq 0$, a comparison is made between this distribution and the corresponding simulated distribution, using the confidence band technique furnished by the Kolmogorov-Smirnov test. The error in the simulated distribution for $\rho_1\rho_2 \neq 0$ is assumed to be $\pm .02$. It is estimated that the error in the approximate distribution for values of $|\rho_1\rho_2| \leq .5$ and $n \leq 30$ is less than $\pm .04$ with 99% probability. For high values of $|\rho_1\rho_2|$ and small n (n = 30) the approximate distribution indicates a larger variance than the true

distribution, hence, resulting in an error that could exceed .04. The results of tests performed on the approximate distribution seem to offer some evidence that for high values of $|\rho_1\rho_2|$ the approximate distribution is accurate only for large values of n ($n > 30$). However, to be able to draw more definite conclusions on the behaviour of the approximate distribution with respect to n and $\rho_1\rho_2$ more thorough testing would have to be performed on the distribution for a wider range of values of the parameters.

The confidence band technique is also used to establish a minimum sample size n with which the approximate distribution for a particular value of $\rho_1\rho_2$ may be applied with reasonable accuracy.

Using Bahadur's theory on sample quantiles, an error bound, ε, can be estimated for each q-critical value, $r_{A,q}(n, \rho_1\rho_2)$, of the approximate distribution for values of $\rho_1\rho_2 \neq 0$. The true critical value, $r_{T,q}(n, \rho_1\rho_2)$, is then given by,

$$r_{T,q}(n, \rho_1\rho_2) = r_{A,q}(n, \rho_1\rho_2) \pm \varepsilon .$$

This paper serves to illustrate how simulation may be used to study a theoretical approximation to an unknown distribution. Simulating realizations of the true unknown distribution of r_{XY} has added a large amount of information on the 'goodness' of the approximate distribution developed by McGregor and Bielenstein

[30]. In studies where data of a particular class is not available, or where analytical or mathematical methods cannot be applied, simulation is perhaps the best approach to solving the problem. In practice, simulation methods have found wide applications only on powerful computing machines. The properties of these methods make them peculiarly suitable for realization on digital computers. Usually simulation methods are also highly dependent on the availability of a 'good' random number generator for generating a large quantity of random numbers on the computer. Hence, they are expensive in terms of computer cost since random number generation by computer is costly, see Table C. Validation of the simulation results also demands careful choice and development of methods of testing the results to determine their representational value and usefulness, and of estimating the error in simulation.

In this thesis simulation was used to obtain more detailed information on the distribution of the sample cross-correlation r_{XY} between two stationary, linear Markov series for both the true unknown distribution and the approximate density function given in [3].

6.3 Further Research.

In our study of the cross-correlation distribution, we obtained, using a sample of 7000 simulated values of r, an error of 2% over the entire simulated distribution. To obtain a reduction in this error would require a significant increase in the sample size N of r and hence the amount of computation. The rate of convergence of the error in the ordinary simulation method is not high and is dependent on $1/N^{1/2}$ (see Shreider [34] and Table D1). Hence, to obtain an error of, say 1% over the entire simulated distribution a sample size > 25,000 would have to be used. This approach is expensive in terms of computer cost and it is obvious that very significant improvement of the accuracy cannot be gained by this means. This observation and the fact that the high computer cost involved in simulation has restricted our testing of the approximate distribution to only a few values of the parameters in Section 5.2, are quite indicative of the inefficiency and impracticability of simple simulation methods as a means of solving distribution theory problems when the entire distribution is required with great accuracy. The ordinary simulation method cannot give any solution of very high accuracy unless special techniques are employed to improve the results. Therefore alternative approaches to improve the accuracy of simulation and modifications of computer simulation methods to increase

efficiency are worth exploring. Several methods, commonly known as variance-reduction techniques, have been proposed for reducing the sampling variability of simulation to increase efficiency. Variance reduction techniques are important in simulation studies since they enable one to improve the accuracy of estimates without increasing the number of tests (or simulation runs). They are, in general, directed towards altering the probability structure of the simulation model so that efficiencies in computation are obtained. Some variance reduction techniques which have been suggested are :

1) Stratified sampling
2) Systematic sampling
3) Importance sampling
4) Correlated sampling
5) Regression methods
6) Use of expected values
7) Russian roulette and splitting
8) Orthonormal functions
9) Antithetic variates
10) Control and Concomitant variates

Kahn and Hammersley [13,19,20] offer two rather complete discussions of variance-reduction techniques. A number of the above-mentioned methods was described and discussed in a recent paper by Gaver [11].

It may be possible and worthwhile in a future research to develop some of the variance-reduction techniques for

application in the present simulation study to obtain more reliable and accurate results, based on which some concrete judgement may be formed on the over-all validity of the approximate distribution.

Our study shows that comparison of the simulated distribution with the true or approximate distribution by the Kolmogorov-Smirnov test is expensive on computer time since the test procedure is laborious, involving at each generation of the r value, the non-decreasing ordering of the r sequence and numerical integration of the $p^*(r)$ function. Furthermore, this test measures the deviation around the region of maximum discrepancy only and is not sensitive to discrepancies in the important tail regions, which may be small but are significant. This indicates the need for more efficient and reliable methods of comparing and testing simulation results, and hence the scope for further research.

While the approximate distribution of cross-correlation may be used to test for correlation between autocorrelated series, its application is limited since it is completely dependent on knowledge of the true autocorrelations which are usually not obtainable in practice. Furthermore, the results of our simulation study seem to indicate that for high autocorrelations of same or opposite sign the approximate distribution is accurate only for large n (that is, longer series). This imposes another restriction on the applicability of the approximate

distribution since in practice series of data available for testing are usually small. Hence, it would be useful if some distribution of the cross-correlation coefficient could be developed in the future for application with estimated values of the autocorrelations and small samples of data even when the absolute value of the serial correlation product is greater than .5.

Other methods of studying the cross-correlation between two series are also worth exploring. Spectral analysis is one approach that may be taken. Since the autocorrelation function and the spectrum (and the cross-correlation and the cross-spectrum) are Fourier transforms of each other, they are mathematically equivalent and therefore have equal representational value. Hence, the spectrum may be used in place of the autocorrelation as a tool in building simulation models and spectral analysis of the results be applied. However, the choice between the spectrum and the autocorrelation as a tool in model building depends upon the nature of the models which turn out to be more useful and efficient. It would be worthwhile to conduct further research into the possibility of applying spectral analysis to multivariate time series in order to derive a distribution of the cross-correlation which is less resticted in application than the approximate distribution that is being studied.

APPENDIX A

ALGORITHM FOR COMPUTING CRITICAL POINTS FOR SAMPLE CROSS-CORRELATION

A1 Evaluation of $p^*(r;n,\rho_1\rho_2)$.

Since the approximate density function of r_{XY}, $p^*(r)$, is an even function of r, it is symmetrical about r = 0. Hence, only positive (or upper) critical points for r_{XY} need to be considered.

Let $p^*(r;n,\rho_1\rho_2)$ denote the approximate density function of r_{XY} for a particular set of values of the parameters $(n,\rho_1\rho_2)$. Let $r_{\alpha;n,\rho_1\rho_2}$ be the upper critical point for r_{XY} at the α- level of significance for that particular set of parameters. $r_{\alpha;n,\rho_1\rho_2}$ may be determined from the relation

$$\int_0^{r_{\alpha;n,\rho_1\rho_2}} p^*(r;n,\rho_1\rho_2)\,dr = \frac{1}{2} - \alpha \quad . \qquad (A.1)$$

Computation of critical points by the above equation involves the numerical integration and evaluation of $p^*(r;n,\rho_1\rho_2)$ for ranges of r-values. A Fortran program is designed to accomplish these two tasks. For numerical integration, the Guass-Legendre quadrature formula is employed. A detailed description of this method of integration is given in Krylov [24] and Shroud [35].

The evaluation of $p^*(r;n,\rho_1\rho_2)$ by computer programming presents both overflow and underflow problems in various terms of the expression in Eqn. (1.5). The following section will discuss the difficulties encountered and methods of avoiding them.

Consider the following terms in the expression of Eqn. (1.5) for the approximate density function, $p(r;n,\rho_1\rho_2)$:

$$K = \frac{2^{M-3}\,(1-\rho_1\rho_2)^{1/2}}{B[\frac{1}{2}M-1\,,\,\frac{1}{2}]} \tag{A.2}$$

and

$$D = \frac{K(1-r^2)^{(M-4)/2}}{[C+(1-\rho_1\rho_2)]^{M-5/2}} \tag{A.3}$$

where

$$B[\tfrac{1}{2}M-1\,,\,\tfrac{1}{2}] = \frac{\Gamma(\frac{1}{2}M-1)\,\Gamma(1/2)}{\Gamma(\frac{1}{2}M-\frac{1}{2})}\,, \tag{A.4}$$

$$C = [(1+\rho_1\rho_2)^2 - 4\rho_1\rho_2 r^2]^{1/2}\,, \tag{A.5}$$

$$M = n + \frac{\rho_1\rho_2(6-5\rho_1\rho_2)}{1-(\rho_1\rho_2)^2} \tag{A.6}$$

Then Eqn. (1.5) may be written as

$$p^{**}(r;n,\rho_1\rho_2) = D[C + (1 + \rho_1\rho_2)]^{1/2} / C \quad . \quad (A.7)$$

From Eqn. (A.6) it can be seen that M is large for large values of n with absolute values of close to zero. That is,

$$M \to \infty \qquad \text{as} \quad n \to \infty \ , \quad |\rho_1\rho_2| \to 0 \ .$$

For small values of n , M becomes negative for values of close to -1. That is, for small n

$$M \to -\infty \qquad \text{as} \quad \rho_1\rho_2 \to -1 \ .$$

Fig. A1 shows the variation of M with values of n and $\rho_1\rho_2$.

However, the restrictions,

$M > 2$ for $(M - 1)$ to have a positive argument

and

$n \geq 6$ for the exponent of $(1 - r^2)$ in Eqn. (1.4) to be positive,

impose a lower bound on the possible values of $\rho_1\rho_2$ for small sample sizes. For $n = 6$,

$$M = 6 + \rho_1\rho_2(6 - 5\rho_1\rho_2) / [1 - (\rho_1\rho_2)^2] > 2$$

requires that

$$\rho_1\rho_2 > -.42 \quad . \quad (A.8)$$

Hence, for small sample sizes, choice of the parameters ρ_1 , ρ_2 must be subject to the condition (A.8).

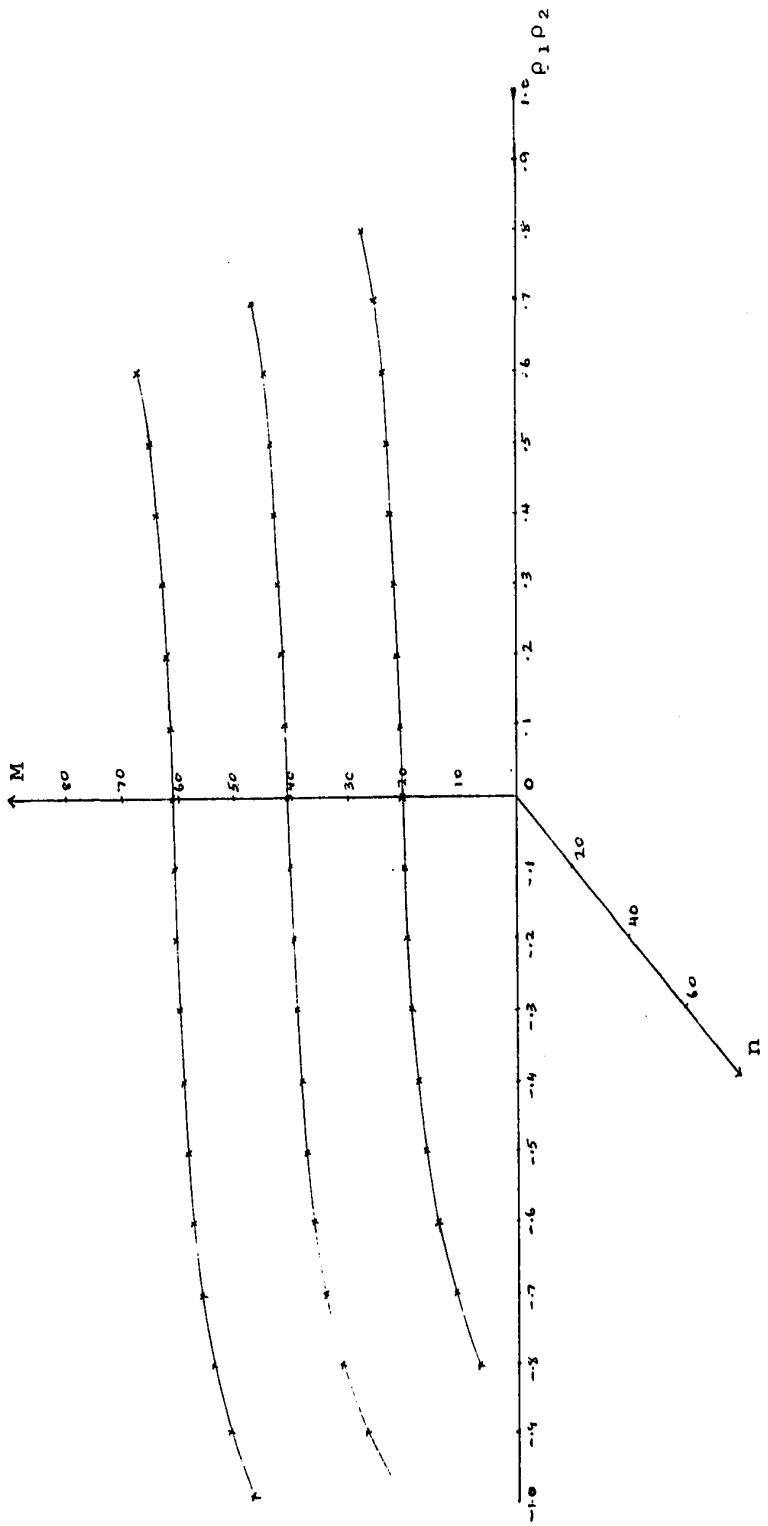


Fig.A1

Values of M against $(n, \rho_1\rho_2)$

In K (A.2), both the BETA function and the factor 2^{M-3} produce overflows for large values of M. To compute the BETA function, Eqn.(A.4) is used. For $M \geq 115$, overflow condition will result in the computation of the GAMMA function since for arguments greater than 57, values of the GAMMA function exceed 10^{75}. In order to handle large values of M, the following Triplication Formula [14] is used

$$\Gamma(3z) = 3^{3z-\frac{1}{2}} \Gamma(z) \Gamma(z + \tfrac{1}{2}) \Gamma(z + \tfrac{2}{3}) / 2\pi \quad . \quad (A.9)$$

This formula can handle values of M as large as 343. For larger values of M, the following Guass Multiplication Formula [14] may be used :

$$\Gamma(kz) = (2\pi)^{(1-k)/2} k^{kz-\frac{1}{2}} \prod_{i=0}^{n-1} (z + i/k) \quad . \quad (A.10)$$

For $M > 130$, the factor 2^{M-3} in K will produce overflows. To avoid this, the k^{th} root (such as k = 3) of K is first computed by

$$K^{1/k} = 2^{(M-3)/k} / \{B[\tfrac{1}{2} M - 1 , \tfrac{1}{2}]\}^{1/k} \quad (A.11)$$

and the resultant $K^{1/k}$ value is then raised to the power k at an appropriate stage of the calculation.

In D (A.3), the term $(1 - r^2)^{(M-4)/2}$ may produce underflows since it approaches zero rapidly as r tends to one. To prevent underflows, the d^{th} root (such as d = 50) of D is computed by

$$D^{1/d} = K^{1/d}(1 - r^2)^{(M-4d)/2d} [C + (1 - \rho_1\rho_2)]^{(2M-d)/2d} \quad (A.12)$$

The resultant $D^{1/d}$ value is then checked against a specified small value ε. If $D^{1/d} < \varepsilon$, D is set equal to zero, and the value of the integral in Eqn. (A.1) is not changed. For example, for $d = 50$, ε is taken to be 0.5 since $(0.5)^{50} \simeq 10^{-15}$.

Taking into account the above considerations, an algorithm and Fortran program may now be developed to evaluate $p^*(r;n, \rho_1\rho_2)$ for a given set of parameters $(n, \rho_1\rho_2)$. The following steps comprise a general procedure for evaluating $p^*(r;n, \rho_1\rho_2)$ at a given r-value :

1) Compute the value of M for the given set of parameters $(n, \rho_1\rho_2)$ using Eqn. (A.6).

2) If $M \leq 2$, $\rho_1\rho_2$ is incremented by δ (that is, $\rho_1\rho_2 = \rho_1\rho_2 + \delta$, where δ is a small increment such as .01) until M becomes greater than 2. This step ensures the proper choice of $\rho_1\rho_2$ for the particular small sample size n used.

3) Compute $B[\frac{1}{2}M - 1 , \frac{1}{2}]$ using Eqn. (A.4) and formula (A.9) or (A.10) for the GAMMA functions (depending on the magnitude of M).

4) Compute $K^{1/k}$ by Eqn. (A.11) , (using, for example, k = 3).

5) Raise the value $K^{1/k}$ to the power k.

6) Compute $D^{1/d}$ by Eqn. (A.12) , (using, for example, d = 50).

7) If $D^{1/d} \leq \varepsilon$ (using $\varepsilon = .5$ for d = 50), set D equal to zero.
If $D^{1/d} > \varepsilon$, raise its value to the power d.

8) Compute C by Eqn. (A.5).

9) Evaluate $p^{**}(r;n,\rho_1\rho_2)$ using Eqn. (A.7),
$(p^{**}(r;n,\rho_1\rho_2 = p^{*}(r;n,\rho_1\rho_2))$.

To facilitate the handling of very large numbers and at the same time achieve greater accuracy in the computation of $p^{*}(r;n,\rho_1\rho_2)$, double precision arithmetic is used to program the above algorithm.

A2 Procedure for Evaluation of Critical Values of r_{XY}.

Using the method developed in Section A1 for evaluating $p^*(r;n,\rho_1\rho_2)$ and Guass-Legendre quadrature formula for numerical integration, an algorithm is deviced to compute the critical points of r_{XY} for appropriate sets of the parameters $(n,\rho_1\rho_2)$. The method of interval bisection is used to estimate the critical points to a desired accuracy.

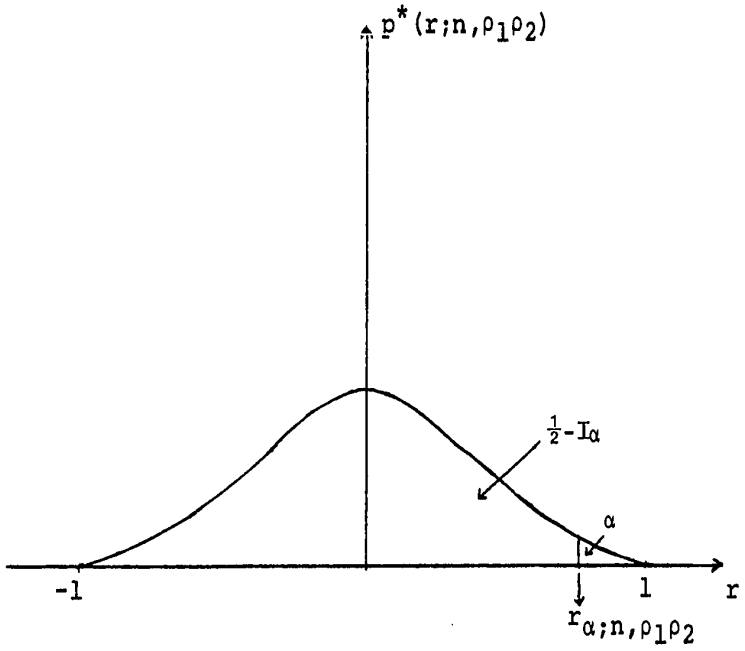


Fig. A2 : $p^*(r;n,\rho_1\rho_2)$ Distribution and Critical Point.

Let I_α denote the area under the $p^*(r;n,\rho_1\rho_2)$ distribution curve as indicated in Fig. A2. Then I_α is given by,

$$I_\alpha = \int_0^{r_{\alpha;n,\rho_1\rho_2}} p^*(r;n,\rho_1\rho_2)\, dr = 1/2 - \alpha \quad . \qquad (A.13)$$

For any point $r_i \in (0\ ,\ 1)$, let

$$I_i = \int_0^{r_i} p^*(r;n,\rho_1\rho_2)\, dr \quad . \qquad (A.14)$$

The algorithm for computing $r_{\alpha;n,\rho_1\rho_2}$ proceeds as follows :

1) Set $a = 0$, $b = 1$, $i = 1$

   Compute

   $$r_i = (a + b) / 2 .$$

2) Compute the area I_i by Eqn.(A.14) using Guass-Legendre quadrature formula for integration.

3) If $| I_i - I_\alpha | < \varepsilon$ for a prescribed ε , r_i is taken to be the value of r_α . [It is possible to arrive at this inequality since the Guass-Legendre quadrature formula used to compute I_i and I_α is known to converge for the $p^*(r;n,\rho_1\rho_2)$ function (see Shroud [35] and Krylov [24]).] Otherwise, go to step (4).

4) If $I_i - I_\alpha > 0$, set $b = r_i$, $i = i + 1$ and go to

step (1) to compute the next approximation.

If $I_i - I_\alpha < 0$, set $a = r_i$, $i = i + 1$ and go to step (1) to compute the next approximation.

The critical points computed by the above algorithm will lie within $\pm \varepsilon$ of the exact critical values for the approximate distribution.

APPENDIX B

JUSTIFICATION OF THE ASSUMPTION THAT THE ERROR IN THE SIMULATED DISTRIBUTION FOR $\rho_1\rho_2 \neq 0$ IS THE SAME AS THE ERROR OBTAINED FOR THE CASE WHERE $\rho_1\rho_2 = 0$.

We note from the plots of the approximate distribution $p^*(r)$ in Fig. 1.1 that the distributions for the various values of $\rho_1\rho_2$ are ' similar in form ' to the normal $N(0, \sigma)$ distribution with zero mean and variance σ^2. For the case $\rho_1\rho_2 = 0$, $p(r)$ is approximately normal for large n. Hence, let us use the normal distributions to validify our assumption concerning the error estimate made in Section 5.2.

Let us consider the $N(0,1)$ distribution as analogous to the $p(r)$ distribution for $\rho_1\rho_2 = 0$. Similarly, let the $N(0, \sigma_1)$ and $N(0, \sigma_2)$ distributions, where $\sigma_1 > 1$ and $\sigma_2 < 1$, be considered analogous to the $p(r)$ distributions for $\rho_1\rho_2 > 0$ and $\rho_1\rho_2 < 0$, respectively. Making an arbitrary choice of the values of σ_1 and σ_2, we simulated the $N(0,1)$, $N(0,.5)$ and $N(0,1.5)$ distributions, using Chen's random number generator and a sample size of 7,000 in each case. We repeated this set of simulations twice, using in each case a different pair of seeds to generate the sample of random numbers. For each of the distributions simulated we computed the corresponding Kolmogorov-Smirnov statistic $D(N)$. If our assumption is valid

the values of D(N) for all nine distributions should be less than $D_{.01}(7000) = .0172$. The results of the simulations are shown in Table B below.

Table B

Kolmogorov-Smirnov Statistics for Simulated Normal Distributions.

N = 7000

Pair of Seeds	748,511,649 147,303,541	281,879,585 27,530,613	983,246,497 858,619,509
Normal Dist.	D(N)	D(N)	D(N)
N(0,.5)	.0063	.0059	.0038
N(0,1)	.0064	.0058	.0036
N(0,1.5)	.0061	.0058	.0039

As shown in the above table, the values of D(N) for all nine distributions are less than .0172. Hence, we can conclude with 99% certainty that the error in all nine simulated distributions is less than .02. This serves to justify our assumption that the error estimate for the simulated distribution with $\rho_1\rho_2 = 0$ holds also for cases where $\rho_1\rho_2 \neq 0$.

APPENDIX C

COMPUTER CONSIDERATIONS IN SIMULATION.

Simulation programs are usually written in a high level language. Fortran IV is used in this simulation study since it is a readily available general-purpose, problem-oriented language most suited to the nature of the problems encountered here. The normal random number generator suggested by Chen is written in Fortran and was found to perform reasonably well on the IBM 360/67.

The major costs of computation involved in this study result from the following :

1) Generation of a large number of normal random numbers in each simulation of the cross-correlation distribution. Each simulation with 7,000 sample values requires over 200,000 random numbers. Approximately 3 minutes are required for each simulation run on the IBM 360/67, using Chen's random number generator.

2) Testing the goodness-of-fit of the simulated distribution by the Kolmogorov-Smirnov test. This test involves, at each generation of the r value, the non-decreasing ordering of the r sequence to compute the empirical cdf, and the

numerical integration of the function $p^*(r;n,\rho_1\rho_2)$ to obtain the theoretical cdf. The numerical integration in the test is done by means of the 32-point Guass-Legendre formula which integrates functions up to degree 63 exactly. A complete test on the simulated distribution by the Kolmogorov-Smirnov criterion, using a sample size of 7000, required about 10 minutes of computer time. It is possible to reduce the amount of computation in the integration by employing a quadrature formula of low order (that is, with less number of base-points). However, this would require knowledge of the degree of the $p^*(r;n,\rho_1\rho_2)$ function to be integrated before the formula can be applied appropriately. Using the 32-point formula this requirement can be avoided since the formula can handle functions of high degrees. Furthermore, for the range of values of the parameters, n, $\rho_1\rho_2$, considered, the degree of the $p^*(r;n,\rho_1\rho_2)$ function is not likely to exceed 63. To improve the accuracy in computation, double precision arithmetic is used to program the evaluation and integration of $p^*(r;n,\rho_1\rho_2)$.

Due to the high cost involved in simulating and testing a distribution with a large sample, the error analysis for the approximate distribution and its critical values could be performed only for a limited number of values of the parameters

$(n, \rho_1 \rho_2)$. Table C below shows the computer timing obtained for some runs of the simulation and testing programs.

Table C

Computer Time Statistics.

(CPU Time Used)

Type of Run	Approx. Time Used (secs.)
Generation of 40,000 normal random numbers.	25
Generation and testing (by Kolmogorov-Smirnov test) of 7,000 normal random numbers.	375
Simulation of cross-correlation distribution for $\rho_1\rho_2 = .72$, $n = 30$, $N = 7000$	160
Simulation and testing (by Kolmogorov-Smirnov test) of distribution for $\rho_1\rho_2 = .72$, $n = 30$, $N = 7000$	769
$\rho_1\rho_2 = .10$, $n = 30$, $N = 7000$	762

APPENDIX D

Table D1

Critical Values $D_\alpha(N)$ for the Kolmogorov-Smirnov Test such that

$$P\{ \max_r |F_T(r;n,\rho_1\rho_2) - F_S(r;n,\rho_1\rho_2)| > D_\alpha(N) \} = \alpha$$

Sample Size N	Significance Level α		
	.01	.05	.10
5	0.669	0.565	0.510
6	0.618	0.521	0.470
7	0.577	0.486	0.438
8	0.543	0.457	0.411
9	0.514	0.432	0.388
10	0.490	0.410	0.368
11	0.468	0.391	0.352
12	0.450	0.375	0.338
13	0.433	0.361	0.325
14	0.418	0.349	0.314
15	0.404	0.338	0.304
16	0.392	0.328	0.295
17	0.381	0.318	0.286
18	0.371	0.309	0.278
19	0.363	0.301	0.272
20	0.356	0.294	0.264
25	0.320	0.270	0.240
30	0.260	0.240	0.220
35	0.270	0.230	0.210
> 35	$1.63N^{-1/2}$	$1.36N^{-1/2}$	$1.22N^{-1/2}$

Adapted from Massey,F.J.,Jr., [29]

Table D2

Critical Values of Cross-Correlation Coefficient, r_{XY} .

$\rho_1\rho_2$ \ n	10			15			20			30		
$\rho_1\rho_2$ \ α	.01	.02	.05	.01	.02	.05	.01	.02	.05	.01	.02	.05
.5	.845	.801	.711	.770	.718	.623	.710	.656	.560	.622	.568	.476
.2	.757	.701	.600	.653	.597	.500	.582	.528	.438	.489	.440	.361
.1	.734	.676	.572	.621	.564	.467	.547	.494	.406	.454	.407	.332
.0	.715	.655	.549	.592	.535	.441	.516	.463	.379	.423	.377	.306
-.1	.704	.640	.532	.567	.509	.417	.487	.435	.353	.393	.350	.283
-.2	.703	.636	.524	.547	.488	.396	.461	.410	.331	.366	.325	.261
-.5						389	.422	.368	.289	.302	.264	.209

BIBLIOGRAPHY

[1] Anderson,T.W. and Darling,D.A., (1954), 'A Test of Goodness-of-Fit.'
J.A.S.A., Vol. 49, 765 - 769.

[2] Bahadur,R.R., (1966), 'A Note on Quantiles in Large Samples.'
Annals of Math. Stat., Vol. 37, 577 - 580.

[3] Bielenstein,U.M., (1963), 'The Approximate Distribution of the Correlation between Two Stationary Linear Markov Series with Fitted Means.'
M.Sc. Thesis (unpublished), University of Alberta.

[4] Birnbaum,Z.W., (1952), 'Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size.'
J.A.S.A., Vol. 47, 425 - 441.

[5] Box,G.F.P. and Jenkins,G.M., (1970), ' Time Series Analysis forecasting and control. '
Holden-Day.

[6] Box,G.F.P. and Muller,M.E., (1958), 'A Note on the Generation of Random Normal Deviates.'

Annals of Math. Stat., Vol. 29, 610 - 611.

[7] Chen,E.H., (1971), 'A Random Normal Number Generator for 32-Bit-Word Computers.'
J.A.S.A., Vol. 66, 400 - 403.

[8] Coveyou,R.R. and MacPherson,R.D., (1967), 'Fourier Analysis of Uniform Random Number Generators.'
Journal of ACM, Vol. 14, No. 1, 100 - 119.

[9] Downham,D.Y. and Roberts,F.D.K., (1967), 'Multiplicative Congruential Pseudo-Random Number Generators.'
Computer Journal, 10, No. 1, 74 - 77.

[10] Feller,W., (1948), 'On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions.'
Annals of Math. Stat., Vol. 19-20, 177 - 189.

[11] Gaver,D.P.Jr., (1972), 'Statistical Methods for Improving Simulation Efficiency.'

[12] Gorenstein,S., (1966), 'Another Pseudo-Random Number Generator.'
Comm. ACM, 9, 10, 711

[13] Hammersley,J.M. and Handscomb,D.C., (1964), ' Monte Carlo

Methods. '
Wiley and Sons, N.Y.

[14] Handbook of Mathematical Functions , (1966), National Bureau of Standards, U.S. Department of Commerce.

[15] Hutchinson,D.W., (1966), 'A New Uniform Pseudo-Random Number Generator.'
Comm. ACM, 9, 6, 432 - 433.

[16] Jannsen,C.T.L., (1970), 'The Management of Bank Reserves.'
(Unpublished), University of Alberta.

[17] Jannson,B., (1966), ' Random Number Generators. '
Victor Petterson, Stockholm.

[18] Jenkins,G.M. and Watts,D.G., (1968), ' Spectral Analysis and its Applications. '
Holden-Day.

[19] Kahn,H. and Marshall,A.W., (1953), 'Methods of Reducing Sample Size in Monte Carlo Computations.'
Operations Research, 1, 5

[20] Kahn,H., (1956), 'Use of Different Monte Carlo Sampling

Techniques.'
Symposium on Monte Carlo Methods
Wiley and Sons, N.Y.

[21] Keeping,E.S., (1962), ' Introduction to Statistical Inference. '
Van Nostrand.

[22] Kolmogorov,A.N., (1933), 'Sulla Determinazione Empirica di une legge di Distribuzione.'
Giorn. Dell'Istit. Degli att., Vol. 4, 461 - 463

[23] Kronmal,R., (1964), 'Evaluation of a Pseudo-Random Normal Number Generator.'
Journal of ACM, 11, 357 - 363.

[24] Krylov,V.J., (1962), ' Approximate Calculation of Integrals. '
Macmillan, N.Y./London.

[25] Lewis,P.A.W., 'The Anderson-Darling Statistic.'

[26] Lewis,P.A.W., Goodman,A.S. and Miller,J.M., (1969), ' A Pseudo-Random Number Generator for the System/360.'
IBM Systems Journal, 8, No. 2, 136 - 146.

[27] Marsaglia,G. and Bray,T.A., (1968), 'On-Line Random Number Generators and Their Use in Combinations.' Comm. ACM, 11, 11, 757 - 759.

[28] Marsaglia,G. and MacLaren,M.D., (1965), 'Uniform Random Number Generator.' Journal of ACM, 12, 83 - 89.

[29] Massey,F.J.Jr., (1951), 'The Kolmogorov-Smirnov Test for Goodness-of-fit.' J.A.S.A., Vol. 46, 68 - 78.

[30] McGregor,J.R. and Bielenstein,U.M., (1965), 'The Approximate Distribution of the Correlation between Two Stationary Linear Markov Series II.' Biometrika, Vol. 52, No. 1-2, 301 - 302.

[31] Miller,L.H., (1956), 'Table of Percentage Points of Kolmogorov Statistics.' J.A.S.A., 51, 111 - 121.

[32] Oak Ridge National Laboratory, (1968), ' Pseudo-Random Number Generator.' University of Alberta Program Library.

[33] Seraphin,D.S., (1969), 'A Fast Random Number Generator

for IBM 360.'
Comm. ACM, 12, 12, 695.

[34] Shreider,Y.A., (1966), ' The Monte Carlo Method. '
Pergamon Press.

[35] Shroud,A.H. and Secrest,D., (1966), ' Guassian Quadrature Formulas. '
Prentice-Hall

[36] Smirnov,N.V., (1944), 'Approximate Laws of Distribution of Random Variables from Empirical Data.'
Uspehi Matem. Nauk, Vol. 10, 179 - 206.

[37] Smirnov,N.V., (1948), 'Table for Estimating the Goodness-of-Fit of Empirical Distributions.'
Annals of Math. Stat., 19, 279 - 287.

[38] Tocher,K.D., (1963), ' The Art of Simulation. '
The English Universities Press, Ltd., London.

[39] Wolfowitz,J. and Levene,H., (1944), 'The Covariance Matrix of Runs Up and Down.'
Annals of Math. Stat., 15, 58 - 69.